DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

Published by Simbios, a National Center for Biomedical Computing

Mining Biomedical Literature:

Using computers to extract knowledge nuggets

PLUS

Successful Collaborations:

Helping biomedicine and computation play well together

Summer 2008

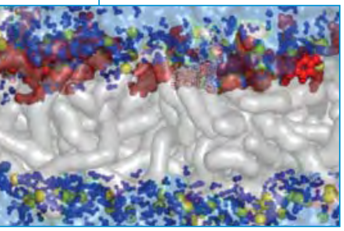
contents

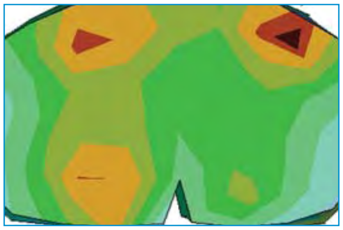
ContentsSummer 2008

FEATURES

DEPARTMENTS

Cover Art
Created by K. Boyack, D. Klavans, and W.B. Paley with data from Thompson ISI. Commissioned by K. Borner and reprinted by permission from Macmillan Publishers Ltd: Nature, 444:985, 2006.





Summer 2008
Volume 4, Issue 3
ISSN 1557-3192

Executive Editor
David Paik, PhD

Managing Editor
Katharine Miller

Associate Editor
Joy Ku, PhD

Science Writers
Katharine Miller
Kristin Sainani, PhD
Meredith Alexander Kunz
Louisa Dalton
Chandra Shekhar, PhD
Roberta Friedman, PhD

Community Contributors
Joy Ku, PhD
John Melonakos
Isaac Kohane, MD PhD

Layout and Design
Affiliated Design

Printing
Advanced Printing

For general inquiries, subscriptions, or letters to the editor, visit our website at www.biomedicalcomputationreview.org

Office
Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

Biomedical Computation Review is published quarterly by Simbios National Center for Biomedical Computing and supported by the National Institutes of Health through the NIH Roadmap for Medical Research Grant U54 GM072970. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. The NIH program and science officers for Simbios are:

Peter Lyster, PhD (NIGMS)
Jennie Larkin, PhD (NHLBI)
Jennifer Couch, PhD (NCI)
Semahat Demir, PhD (NSF)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Richard Morris, PhD (NIAID)
Joseph Pancrazio, PhD (NINDS)
Grace Peng, PhD (NIBIB)
David Thomassen, PhD (DOE)
Ronald J. White, PhD (NASA/USRA)
Jane Ye, PhD (NLM)
Yuan Liu, PhD (NINDS)

BY ISAAC KOHANE, MD, PhD

When Does Computational Validation Trump Biological Validation?

Many a successful investigator working at the interface between molecular biology, genetics and computation will recognize the imperative to obtain biological validation for computational investigations. Even if they have extensively mined multiple datasets of prior research done by others, experience will have shown that the lack of an additional, novel validation dataset will make it challenging to overcome the reviewers' concerns. This is particularly the case for the top tier, "high impact" journals. Of note, this expectation of additional, novel biological validation will be stated not only by reviewers from a traditional molecular biology or genetics background, but also by many of us in the bio-computational community. I wish to argue here that in many instances, such requirements are the result of an inadequate understanding of the nature of the data being used and their value as compared to a novel incremental dataset. Moreover, such requirements represent a failure of the bio-computation community's confidence in their own methodology and a similar failure in our ability to educate our broader biological investigational community regarding what constitutes a figure of merit in a modern computationally-assisted scientific investigation.

I was recently reminded of this failure when I presented results from work involving our National Center for Biomedical Computing, i2b2 (Informatics for Integrating Biology and the Bedside) during a session at a Keystone meeting on insulin resistance. Using multiple datasets from one of my previous collaborations with the Joslin Diabetes Center in Boston, we had undertaken a simple meta-analysis across multiple experiments conducted by leading investigators in type 2 diabetes involving mouse models or human models of insulin resistance. In collaboration with i2b2 investigator **Peter Park, PhD**, we found that no gene was significantly expressed across all the murine and human models. Nonetheless we found that in 8 of 17 experiments one gene was differentially expressed—which I thought remarkable given the diversity of heterogeneous mouse and human experiments involved. Yet after my presentation, colleagues expressed skepticism about the validity and interest of these results, given that the analysis brought together so many disparate conditions and organisms. The dominant scientific culture expects novel results to arise only under a highly specific set of conditions in individual investigator's laboratories. However, **Dr. Mitch Lazar** from the University of Pennsylvania came to the podium immediately after my presentation and generously remarked that I had scooped him! His own research (a genome-wide chromatin immuno-precipitation scan) had also revealed the significance of that same gene in insulin resistance and adipogenesis, a result he confirmed in several *in vitro* studies. Dr. Lazar's results will soon be published in a first tier journal—and deservedly so. Yet it would have been very challenging for our purely computational analysis to receive similar treatment.

There are without a doubt several purely computational analyses from which any biological conclusions drawn are suspect. Further experimentation or data are required before any tentative conclusions can be drawn. Equally suspect, however, but far more often published, are biological results from an *in vitro* experiment in a non-human model organism under conditions having little to do with those experienced in the course of human pathology. Nonetheless there is a class of computational investigations that leverage prior, often published data sets, sometimes singly and sometimes together. Can we establish a scientific theory or at least a reliable set of heuristics as to when such investigations are sufficient? Are there conditions when an overwhelming set of "lightly used" previously published data can be re-explored to even greater effect and greater generality and applicability than a narrow set of biological experiments? Are there indeed a set of computational investigations that require no additional biological validation? Those of us who work at the intersection of computation and biology are both the best placed to provide principled answers to these questions and also should be the most motivated to so. Let the games begin. □

DETAILS

Isaac Kohane, MD, PhD, is Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology at Harvard Medical School; Chair of the Informatics Program at Children's Hospital, Boston; and Principal Investigator for Informatics for Integrating Biology and the Bedside (i2b2) a National Center for Biomedical Computing.

NewsBytes

Why We Swing

Most people swing their arms when they walk. Indeed, like several characters in a classic Seinfeld episode, we're surprised when they don't. Yet we don't really need to swing our arms in order to move forward, as we all know when we carry a box with both hands. So why do we swing our arms when we walk? A recent computational model by **Jaeheung Park, PhD**, a re-searcher at the Stanford Artificial Intelligence Laboratory at Stanford University, provides some insight.

Arm swinging, Park hypothesized, serves the same purpose as rotational friction—the friction between the foot and the ground that keeps our feet from turning in or out like windshield wipers. And his simulations, published in the *Journal of Biomechanics* in April 2008, confirmed that possibility.

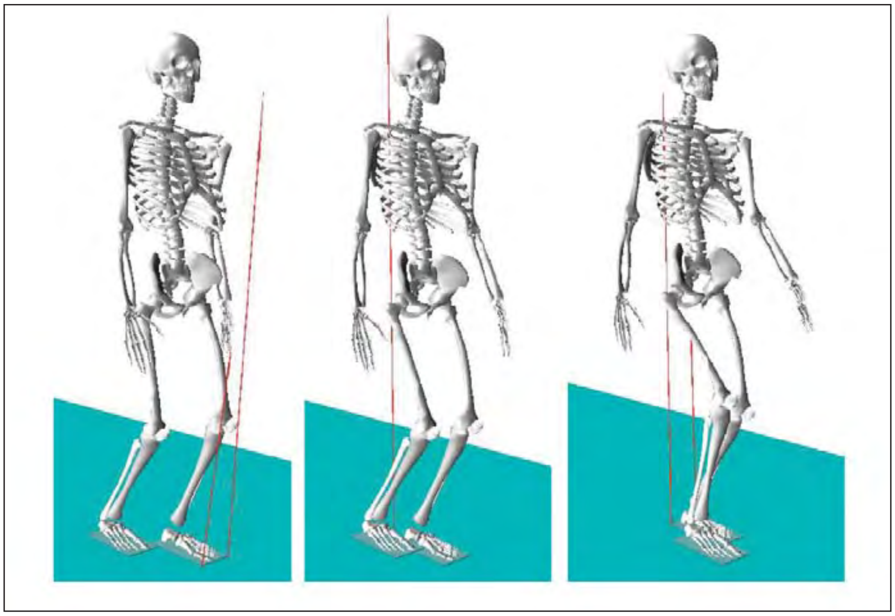
In the past, many biomechanical models of gait have omitted the arms. But as such models strive for greater realism, it has become more important to account for secondary movement by the arms. One way to do that is to simulate the trajectories of the arms and joints. But Park took a different "task-oriented" approach adapted for human simulations from his thesis advisor's work on industrial robots.

In his simulations, Park instructed the feet to perform a task—"walk"—but gave no instructions to the arms. Then he varied the amount of rotational friction between the foot and the ground. When the rotational friction forces experienced by the model's foot were large enough to minimize body movement, the arms didn't swing. They didn't need to. But when the rotational friction at the foot was constrained to zero, the arms swung naturally in compensation. This was true for two different styles of walking—static (a kind of slow stagger where the center of mass is always over one foot or the other or both) and dynamic (a more realistic style at a normal human pace).

To Park, these results suggest that arm swinging helps us maintain our balance on slippery surfaces because it compensates for the absent rotational friction. In addition, it provides greater comfort, since the foot and consequently all the leg joints do less work.

"This paper has elucidated the relationship between arm swing and the support moment at the foot," comments **Marcus Pandy, PhD**, chair of mechanical and biomedical engineering at the University of Melbourne, Australia. More work remains to be done, though, to understand the relationship between the foot's role and "energy consumption during gait." Pandy also notes that "it would be interesting to see how the joint torques predicted by the model compare with those obtained from experiments when humans walk at their preferred normal speeds."

In the future, Park would like to explore whether arm swinging affects the speed of movement. Eventually, such work might provide more evidence that there is a good reason to swing.
—By Meredith A. Kunz



A simulation of human walking with zero friction at the foot generates natural arm swing motion. Courtesy of Jaeheung Park. Reprinted from Journal of Biomechanics 41: 1417-1426, 2008 with permission from Elsevier.

RNA Takes Shape

RNA is not just a single-stranded template. Like proteins, many RNA molecules can fold into three-dimensional structures that catalyze reactions and regulate gene expression. Predicting this structure, though, remains an open challenge. Scientists at the University of Montreal have devised a novel way to attack the problem, which they describe in the March 6 issue of *Nature*.

"Our approach is to generate a more complete RNA secondary structure and from there go to three dimensions directly. Whereas before going to 3-D from secondary structure was impossible," says **François Major, PhD**, professor of computer science and operations research, who developed the method with graduate student **Marc Parisien**.

RNA nucleotides bind with each other to form secondary structures such as hairpins (a stem with a loop) and helices. Though most nucleotides pair according to Watson-Crick or wobble rules (C-G, A-U, and G-U), a small number (about 15 percent of nucleotides in hairpins, for example) form alternate pairings—such as A-C or a G-U-A base triple (where the bases meet in different orientations). Previous programs have fallen short of predicting these "non-canonical" pairings that are the key to

3-D structure and indeed often drive the most interesting geometries such as loops, bulges, and twists.

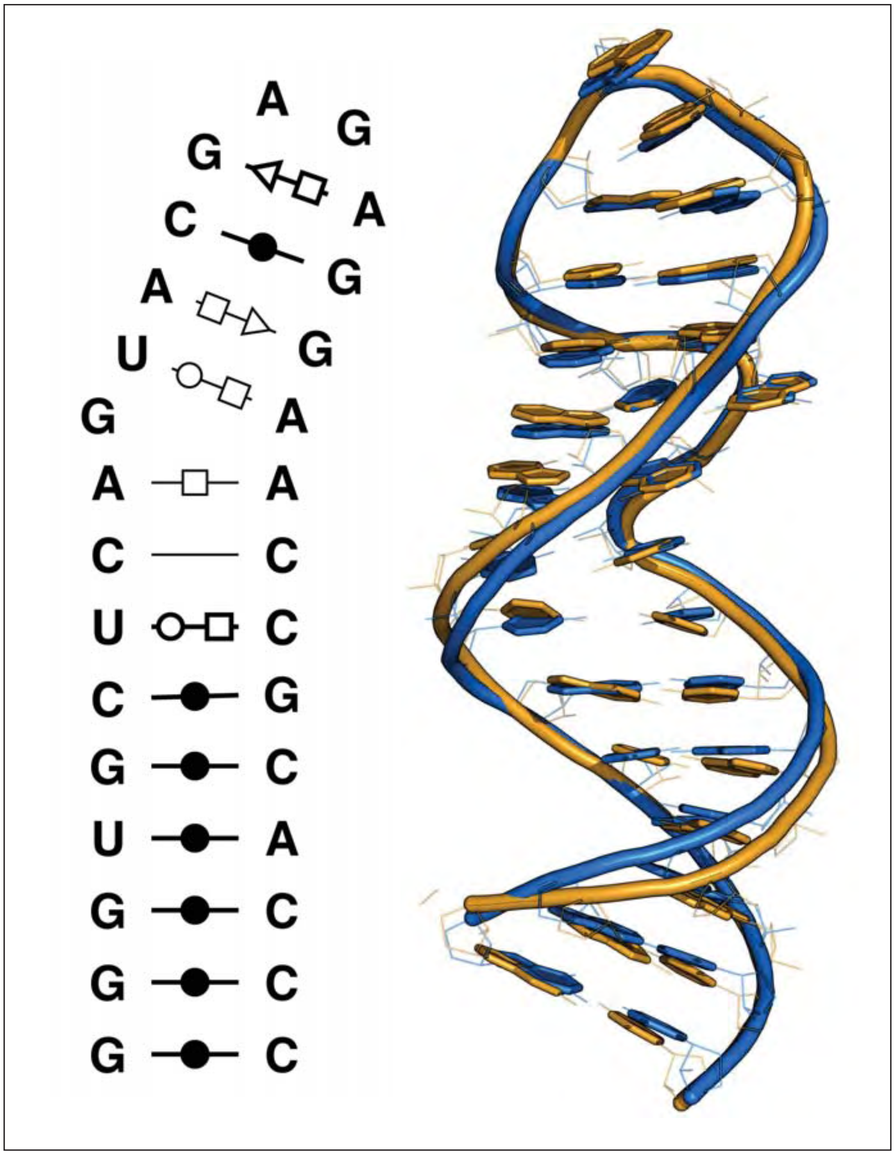
To better predict non-canonical pairings, Major and Parisien identified 19 regular, repeated small motifs (mostly 3 to 5 nucleotides) in solved RNA structures. They call these the RNA structural alphabet or "nucleotide cyclic motifs" (NCMs). The most common "letter" (or NCM) consists of two Watson-Crick base pairs stacked on top of each other; a bunch of these together form a basic helix. But many of the other NCMs are defined by non-Watson-Crick base pairs. One example is a four-nucleotide loop with a G-A pair at the bottom.

To determine the 3-D structure of a given RNA primary sequence, Major and Parisien feed it through two programs: MC-Fold and MC-Sym. MC-Fold enumerates all possible base pairings (including non-canonicals) and all possible arrangements of NCMs. It then picks the most probable arrangement based on statistical data from solved RNA structures. Next, MC-Sym translates the NCMs directly into 3-D structures. The pipeline is available as a web service (http://www.major.iric.ca/MC-Pipeline/). Currently, accuracy is limited to sequences of fewer than 75 base pairs—unless experimental or multiple-sequence data are incorporated into the program, Major says.

As a test case, Major and Parisien folded several precursor microRNAs (with previously unknown structures). Such molecules would be expected to share a common structural element for binding to the enzyme Dicer, which processes them into functional microRNAs. The result: despite different primary sequences as well as non-canonical base pairs and bulges, the pre-microRNAs all folded into double helices.

"That's a pretty powerful result," comments **Philip Bevilacqua, PhD**, professor of chemistry at Penn State University. "I think this method is going to be of practical benefit to the RNA community," he says. "This has the potential for enormous impact, and hopefully it will get fulfilled."

—By Kristin Sainani, PhD

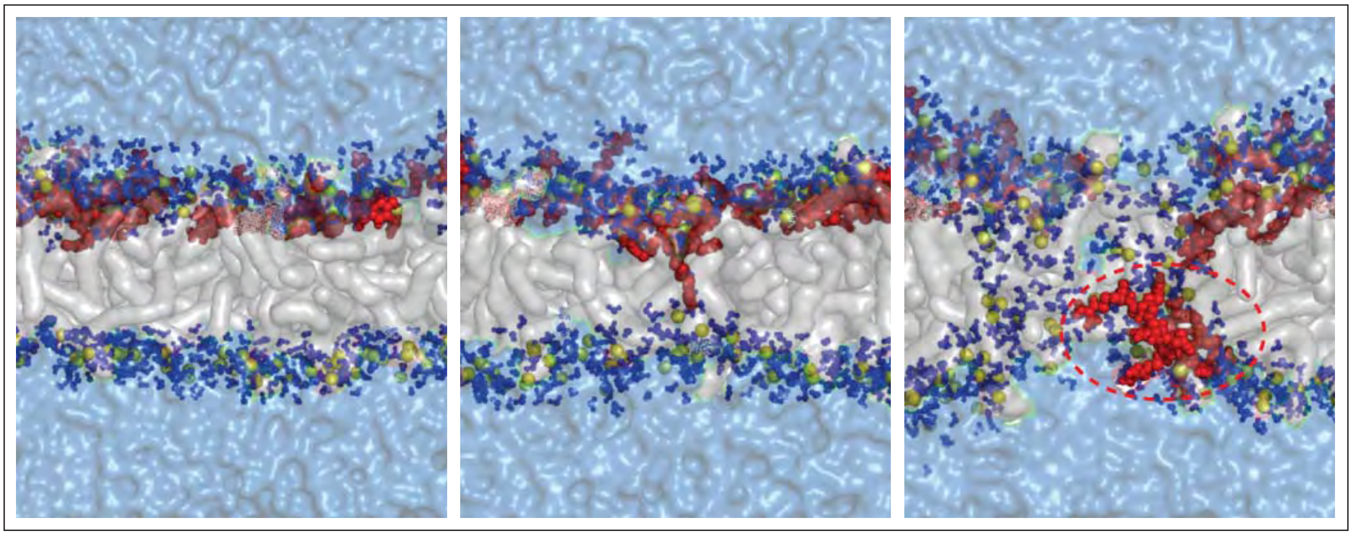


Predicted 2D and 3D structures of an RNA loop (Sarcin/Ricin loop from rat ribosomal RNA). Left: Watson-Crick (black dots) and non-Watson-Crick base pairs predicted by MC-Fold. Right: predicted 3D structure (blue) superimposed on the experimentally determined structure (gold). Courtesy of Francois Major and Mark Parisien.

Trojan Peptide

A powerful snippet of protein called the Tat peptide ferries itself across cell membranes dragging just about anything it's attached to along with it. How it accomplishes this feat has been a puzzle for a decade. Now, computational simulations offer a detailed picture of how the string of eleven amino acids cajoles the membrane's lipid bilayer into doing most of the work.

"I was expecting that the peptide would act like a snake going through a hole," says **Angel Garcia, PhD**, professor of biocomputation and bioinformatics at Rensselaer Polytechnic Institute, who helped design the simulations. Yet his laboratory's simulations suggest that instead of the snake doing all the work, it is as if the ground makes space for the snake to pass. "I wasn't expecting the lipids to change so drastically," he adds.



At left, four Tat peptides (red) cluster on one side of a lipid bilayer (white) attracted to the phosphate groups (yellow). As the Tat peptide reaches toward phosphate groups on the opposite side (middle), the bilayer thins enough for a chain of water molecules (blue) and the peptide to pass through the membrane (right). Courtesy of Angel Garcia. Reprinted from Proceedings of the National Academy of Sciences *104:52 (2007).*

"Once you see it, of course, it could not be any other way." The work was published in *Proceedings of the National Academy of Sciences* in December 2007.

The Tat peptide, discovered on an HIV protein, is part of a potent group of cell-penetrating peptides sometimes called Trojan horse peptides. They haul drugs, proteins or DNA right across the lipid bilayer and into the cell. The myriad uses of such peptides in both therapy and research are not hard to imagine. But how these highly charged, water-loving bits of protein so readily cross the waterless middle of the lipid bilayer has evaded answer for years.

Garcia and postdoctoral fellow **Henry Herce, PhD**, decided to apply the power of a new computer center at RPI to conduct molecular dynamics simulations of the Tat peptide as it approaches and crosses a lipid bilayer.

Over and over again, the simulations reveal how the peptide induces a change in the bilayer. Because six of the eleven amino acids in Tat are arginine, a relatively large, positively charged amino acid, researchers knew that Tat would be strongly attracted to the lipid bilayer with its blanket of negatively charged phosphates. But Garcia did not expect that phosphates on both sides of the bilayer—not just on Tat's side—would align to help neutralize Tat's charge. The more peptides added to the mix, the greater the influence on the opposite side of the bilayer. As the arginine side chains and distant phosphate groups move toward each other, the bilayer thins until it creates a hole lined with phosphate groups, letting a small chain of water and the peptide pass through.

"The idea that the bilayer is 'thinned,' thereby allowing the cationic TAT to touch anionic phosphate head groups on both sides of the membrane was utterly unexpected," says **Steven Dowdy, PhD**, a Howard Hughes investigator and professor of cellular and molecular medicine at the University of California, San Diego. Dowdy says the information from Garcia's computational work will inspire experimental testing of the mechanism. And, he says, it could be very helpful in designing enhanced peptides with increased potential to deliver drugs or DNA where researchers want them.

—By Louisa Dalton

Window into Microbial Behavior

We know they are there, but most microbial denizens of deep oceans, sea floor vents, even our own intestines, remain a mystery. Because most microbes won't grow in the lab, researchers have few clues to their communal activities. With better gene sequencing and computational ability, researchers now sample genes from whole communities to assemble the "metagenome"—a picture of the genes driving metabolic processes important to growth and survival in a given environment.

In a new study, researchers found remarkable diversity in how microbes function in each of nine distinct biomes. Indeed the bacterial and viral genomes from each biome had distinguishing metabolic profiles. And viral genomes—which researchers expected would be similar across environments—were just as different as the bacteria.

It turns out that there's a surprisingly extensive genetics arms race going on between bacteria and the viruses (called phages) that infect them, says **Rob Edwards, PhD**, assistant professor in the Computational Sciences Research Center at San Diego State University. Viruses are actively shuffling their host bacteria's DNA. "We didn't know (just) how much DNA the viruses move around," Edwards says. In fact, it happens so often that, he believes, the viruses likely profit from moving pieces of DNA that are beneficial to the bacteria.

Edwards and his collaborators from San Diego State University, Argonne National Laboratory and around the world reached these conclusions by comparing nearly 15 million sequences from

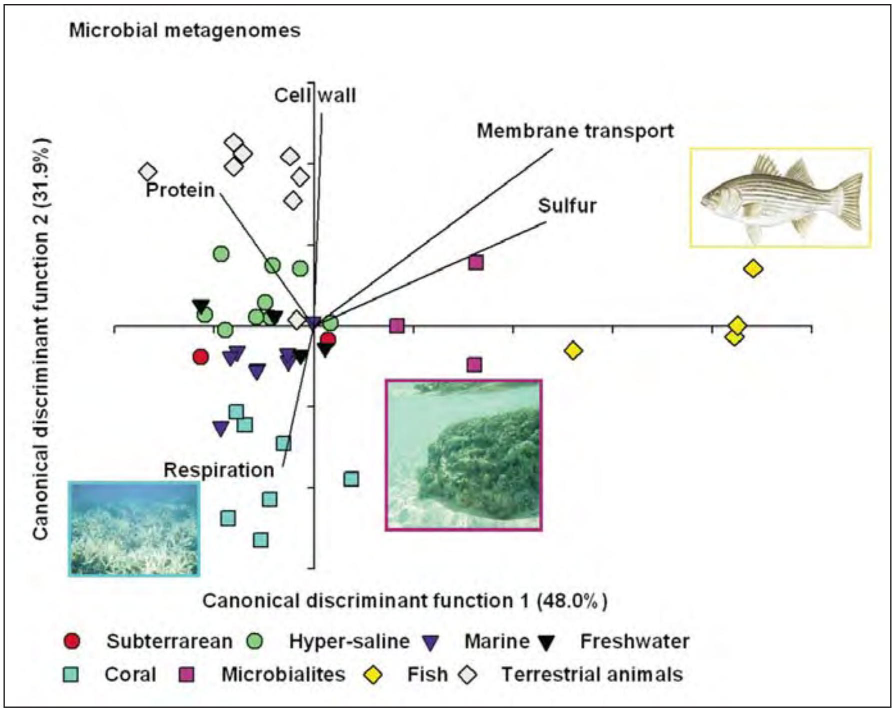
45 microbial communities, including 42 viral genomes, as reported in *Nature* on April 3, 2008. It's easy and relatively inexpensive to generate a DNA sequence these days, Edwards says, "What is not so easy is to figure out what it actually means."

Thanks to the SEED database (www.theseed.org), developed in collaboration with researchers at Argonne Labs and the Fellowship for Interpretation of Genomes, which annotates or assigns known function to gene locations, scientists can upload gene sequence data and seek a pattern of metabolic activities that exist in their samples. They can thereby begin to compile the collective activities of a given community, be it a coral reef, a mine shaft, or a person's bronchi.

This sort of work will definitely help researchers understand and harness the functions of bacteria, says **Eric Delwart, PhD**, a virologist at the Blood Systems Research Institute and the department of Laboratory Medicine at the University of California, San Francisco. "Bacterial genomes are scrambled and slapped together by viruses. The core functions probably cannot be exchanged, but peripheral functions can be passed around," he says, in a process unique to bacteria that likely speeds up their rate of evolution.

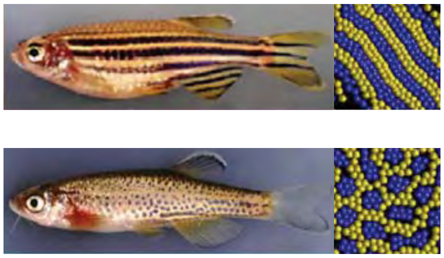
Such gene swapping may also yield therapeutic insight. A lot of diseases, such as atherosclerosis and stomach cancer, have "a very strong microbial component," Edwards says. "We are working with the NIH to get at the bioinformatics of this."

—By Roberta Friedman, PhD



The survival techniques of bacteria in nine different biomes (represented by different colored symbols) can be distinguished based on the prevalence of various metabolic gene subsystems such as respiration, membrane transport, virulence, or sulphur metabolism. The length of the lines represents the degree of influence of a metabolic process. Courtesy of Elizabeth Dinsdale. Reprinted by permission from Macmillan Publishers Ltd: Nature, 452, 629 - 632 (12 Mar 2008). *Coral and microbialite photos by F. Rohwer.*

How the Zebrafish Gets its Stripes (or Spots)



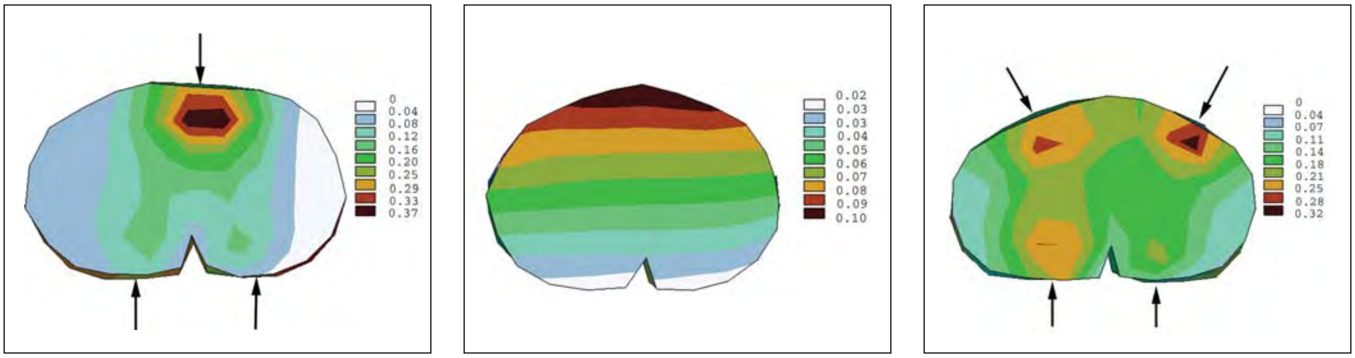
Normal and leo mutant zebrafish (left) and their corresponding normal and abnormal simulated patterns (right). Courtesy of Troy Shinbrot and David Parichy. Reprinted from Developmental Biology, *315, Caicedo-Carvajal, CE; Shinbrot, T, In Silico Zebrafish Pattern Formation, 397–403 (2008), with permission from Elsevier.*

Normal zebrafish have stripes, but mutant forms may display spots, blotches, or labyrinthine patterns. It's a scenario that Rudyard Kipling might turn into a wonderful "just-so" story. But a more scientific explanation comes from a new computer model that can replicate the diverse ways that pigmented cells organize themselves on zebrafish skin. The results may help scientists gain a better understanding of development in general, helping explain how myriads of cells turn into tissues, organs, and entire organisms.

"Our aim here is not to build better zebrafish," says **Troy Shinbrot, PhD**, who developed the model along with his graduate student, **Carlos Caicedo-Carvajal**. "We want to understand how tissues and organs develop and how cells migrate, survive, and form the shapes that govern function." The work was published in *Developmental Biology* in January 2008.

According to Alan Turing's theory from the 1950s, pigmented cells arrange themselves into patterns under the guidance of chemical agents. More recent studies of zebrafish stripe formation suggest that mechanical interactions between cells—how strongly they push or pull one another—could also play a vital role. To test the latter hypothesis, Shinbrot and Caicedo-Carvajal developed a simplified energy-minimization model of cells of two different colors interacting within a rectangular region.

Using different combinations of values for the forces between like (homotypic) and unlike (heterotypic) cells, the researchers generated a range of possible patterns. "To get stripes, we need both heterotypic attraction and a delicately balanced homotypic repulsion," says



These cross-sections of a simulated spinal cord show the different deformation patterns induced when the cord is subjected to a transverse contusion injury (left), a distraction injury (center) and a dislocation injury (right). Courtesy of Carolyn Greaves. Reprinted from Greaves, C, Gadala, M; Oxland, T, A Three-Dimensional Finite Element Model of the Cervical Spine with Spinal Cord: An Investigation of Three Injury Mechanisms,* Journal of Biomechanical Engineering *36:396 (2008) with kind permission of Springer Science and Business Media.

Shinbrot. If these conditions were not met, the simulations showed that spotted, striated, labyrinthine, and other non-striped patterns developed; in particular, when all the inter-cellular forces were attractive, only spots formed. The researchers showed that some of these abnormal patterns resemble those observed on certain mutant zebrafish varieties with defective pigment pathways.

This *in silico* approach could be applied to a broad range of problems in cellular development, says Shinbrot. The researchers are now using it to help oncologists compare four different patterns of abnormal tissue commonly seen in early breast cancer tumors.

"The authors have done a nice job of showing how you can produce a whole repertoire of patterns simply by tuning the strengths of attractive and repulsive cell interactions," says **Ed Munro, PhD**, a computational cell biologist at the University of Washington in Seattle. However, Munro cautions that the results obtained using the authors' simplified model need further biological validation. "By demonstrating one way in which cells can make patterns, you haven't shown that's how embryos do it," he notes.

—*Chandra Shekhar*

Modeling the Spine, Cord and All

When the bones and discs of the spinal column are broken, crushed, or displaced, the spinal cord itself may be devastatingly damaged. Now, a new computer model suggests that the manner in which the injury occurred may affect the spinal cord in distinct and significant ways.

This work could have a wide-reaching impact on spinal treatment, says **Thomas Oxland, PhD**, professor of orthopaedics and mechanical engineering at the International Collaboration on Repair Discoveries (ICORD) Centre at the University of British Columbia. If cord injuries could be subclassified by type, it is possible that physicians may be able to treat them differently. Oxland was lead author of the work, published in *Annals of Biomedical Engineering* in March 2008.

Before modeling the human spine, Oxland's team, which included his master's student **Carolyn Greaves**, and **Mohamed Gadala, PhD**, a professor of mechanical engineering, had already begun animal studies to examine the relationship between the type of spinal column injury and the strain on the cord. But they wanted to compare their animal data to the human spine. Because it's impossible to use human experimental models, the group simulated the spine and the spinal cord using data from the Visible Human project.

Like others who have modeled the spine, Oxland and his colleagues created a finite element model of the human cervical (neck) spine. They then simulated injury to it by applying engineering torques, not unlike those used to study the strain on a bridge. What's new here is that they observed the effect of different types of injuries on the spinal cord itself. The result: distinct patterns of strain and deformation depending on whether the spine suffered a burst fracture, a dislocation, or a stretching injury. The work stopped short of examining actual cord damage but, Oxland says, "one would expect that [these mechanisms] would produce very different patterns of damage in the cord."

Oxland acknowledges that their now-static model cannot yet capture the dynamic forces at work when a real-life injury happens, often in a fraction of a second. His team is working on introducing more variables and lifelike properties now. He also plans to match up the simulation results with his lab's animal experiment data to better understand cord damage.

David Shreiber, PhD, an assistant professor of biomedical engineering at Rutgers University, thinks this model will help advance the field—one that still lags behind brain injury research. "It's significant because it's the foundation of more work on injury to the cord," says Shreiber. The model is flexible enough that it can be used to understand many types of injuries. "The nice thing about this computational system is that you can apply the loading conditions however you want—you can look at twisting, at pressure applied internally, and other cases of spinal injury," he adds.

—*Meredith A. Kunz* □

Successful Collaborations:

Helping biomedicine and computation play well together

BY KATHARINE MILLER

Social scientists who study science have noticed a trend: More and more researchers are collaborating. Over the last twenty years, the number of co-authored papers has increased in every scientific discipline and across diverse geographic areas. Co-authored papers are also cited more frequently than single-authored papers, according to what are called "bibliometric" studies.

And many collaborations bridge disciplines. Biomedical computation—interdisciplinary by nature—is no exception. Many of its goals require the involvement of people with different expertise.

So if collaborations will be a fact of life for many involved in biomedical computing, what can be done to make them productive? Can social scientists provide any insights?

Skepticism abounds about whether social scientists' observations of scientists are more informative than scientists' own experience. The ingredients of a successful collaboration seem obvious: good leadership, trust among the participants, face-to-face meetings and strong communication skills.

But then why do many initial collaborations fail? Studies show that even when collaborators are in the same location (a best case scenario), fewer than a third of collaborations succeed, says **Gary Olson, PhD**, a professor of human computer interaction at the University of Michigan who has studied collaborative science. So if common sense can only take us so far, perhaps rigorous research is needed to fill in the gaps. Sociological research can produce counterintuitive findings; answer debates about contrasting models of collaboration; and provide specifics about what works and what doesn't work. "We have an idea about what factors matter, how they matter, and how to intervene to make a collaboration work," Olson says.

Judy Olson, PhD, also a professor of human computer interaction at the University of Michigan, has developed a "Theory of Remote Scientific Collaboration." The "TORSC," as it's known in social science circles, describes a number of factors that can affect the success of collaborations. As suggested by the word "remote" in the theory's name, the most important factor is distance itself. In addition, collaboration readiness, technical readiness, modularity of tasks, and a management plan can make a huge difference—as the leaders of various collaborations attest below.

Some believe that changes in the next generation's social world—which is so reliant on computer interaction—will alter the collaborative landscape. If so, it will provide plenty of fodder for further sociological study. But for now, using guidelines based on the TORSC factors seems wise.

PROXIMITY MATTERS

Thirty meters. That's the rule of thumb. When coworkers are located more than 30 meters from one another, a collaboration's effectiveness declines precipitously, according to the "Allen Curve" discovered by Massachusetts Institute of Technology researcher **Thomas Allen** in the 1970s. Accidental meetings in the hall, water-cooler conversations, lunchroom chats—all of these unplanned encounters between collaborators drop off beyond that distance. And that matters a great deal, says **Jonathon Cummings, PhD**, associate professor of management at Duke University's Fuqua School of Business. It becomes harder to foster a collegial social environment, build common ground, maintain awareness of what others are doing, attend to the project, and adjust to surprises, research shows.

Collaborators located in different buildings within the same university or institution can compensate for a lack of informal interaction with regular in-person meetings. "But as soon as it's not built into their plans, then the 30-meter rule really does operate," Cummings says.

Furthermore, many collaborations—including biomedical computation collaborations—happen at distances measured in miles rather than meters. They range across differ-

ent states, regions, or countries and connect multiple institutions. And these distances really do get in the way of success, according to many sociological studies.

Research by Cummings and his colleague **Sara Kiesler, PhD**, Hillman Professor of Computer Science and Human-Computer Interaction at Carnegie Mellon University, specifically shows that multi-institutional collaborations are less successful than collaborations within a single institution. They studied nearly 500 collaborations funded under the National Science Foundation's Information Technology Research (ITR) program. The collaborations were relatively small—five to ten principal investigators each. But even at that scale, they found that as the number of universities involved increased, researchers spent less of their funding on practices that foster collaboration. They held fewer meetings, made less effort to divide up responsibilities effectively, and transferred less knowledge (such as the best way to do things) from one part of the organization to another. The result: collaborations involving more institutions actually generated fewer positive outcomes (such as papers, new models, new ideas, new software, spin-off projects, or PhD dissertations for graduate students).

"So in a sense, these distributed projects are shooting themselves in the foot by not investing in the very things that would help them succeed," Cummings says. Cummings believes this happens because of budgetary selfishness. If you're spread across multiple institutions, he says, "you're more likely to spend the money on your own institution's needs—salaries and graduate students—than to spend it on a shared workshop or conference."

This finding raises an important question: as interdisciplinary research collaborations become more and more common, should they be promoted more within universities or between them? In recent years, top universities have invested heavily in bringing multiple disciplines physically together under one roof. The Clark Center at Stanford, the Lewis-Sigler Institute for Integrative Genomics at Princeton, and the Broad Institute of MIT and Harvard, represent a few prominent examples. In contrast, funding agencies have been "looking to get the most bang for the buck" by supporting between-university collaborations, Cummings says. Cummings' research supports the former strategy. If given $100 million to invest in either a Clark Center or a collaboration among multiple institutions, he says he would "without a doubt" build a Clark Center.

But **Mark Ellisman, PhD**, director of the Biomedical Informatics Research Network (BIRN), thinks otherwise. BIRN, launched in 2001, was one of the first large-scale biomedical "collaboratories"—a term that refers to large distributed collaborations that rely heavily on tools of the digital age. BIRN, a National Institutes of Health initiative, consists of 31 research groups at 23 universities around the United States and in England. All are working together on infrastructure development and three projects centered around the imaging of human or mouse brains.

As a result of his experience with BIRN, Ellisman thinks we're coming to the end of the era when universities need to attract the best and brightest to their own faculty. "I can gather the best scientists in the world in a virtual collaboration more quickly, and act to conquer big challenges," he says. "I don't need to find a way to move everybody here."

"I can gather the best scientists in the world in a virtual collaboration more quickly, and act to conquer big challenges," says Mark Ellisman. "I don't need to find a way to move everybody here."

Cummings concedes that it may not be possible to address some complex problems without involving multiple institutions—particularly where there's scarce expertise or scarce equipment. "But I would say those are rare or far less likely than other types of projects."

And Cummings admits that distance isn't everything. As described below, readiness and modularity of tasks can, to some extent, help to overcome the problem of distance. As can taking the lessons of Cummings' research to heart: if you're planning a multi-institutional collaboration, set aside money and time to make it all come together—plan on holding symposia; dividing up tasks effectively, and meeting face-to-face on a regular basis. Don't leave it to chance, or it won't happen—unless your institutions are less than 30 meters apart.

COLLABORATION READINESS

Collaborations collapse when people don't have the right motivations and experience before they launch—what Gary Olson's group calls "collaboration readiness." "All kinds of projects fail when people try to collaborate because the funding agency said they had to or they think they'll get more money if they collaborate—exogenous reasons that don't really make the collaboration work," Olson says.

For example, in the case of the ITR projects that Cummings studied, few of the researchers involved had worked together before. "The funding agencies were holding a big carrot that says we'll only fund you if you put together a distributed interdisciplinary project," Cummings says. "So you had all these people clamoring to find partners to collaborate with. In my view, that's kind of going about it all wrong. . . . Taxpayer money is being spent on sort of trying out a relationship." In a separate study, Cummings showed that people who have never worked together before are less likely to overcome the barriers of working across institutions.

Ellisman admits that in the early days, BIRN researchers had to overcome a lack of collaboration readiness. "Many collaborators were happy to see extra research dollars but actually doing something beyond the office next door, or having to sit in video teleconferences was a bit more painful. They had to work up to it."

And participants have sometimes been unwilling to share data too soon—a factor in collaboration readiness. But, as Ellisman puts it, "With NIH saying 'thou shalt do things differently,' we got everyone to sign off that if they were going to take money under BIRN, there would be open access."

A lack of common ground can also raise readiness concerns, as BIRN researchers discovered in the Mouse BIRN project: Researchers from different subdisciplines referred to the same location in the brain using different terminology. BIRN solved this problem by building ontologies to establish a common vocabulary and by creating a "Smart Atlas," which allowed data to be placed within a common coordinate system.

TECHNICAL READINESS

The TORSC holds that remote collaborations must also exhibit "technical readiness," meaning the participants must be comfortable with the use of communication tools that make long-distance collaboration easier. "If physicians are working together and the video conferencing tools don't work, they'll just walk away," says Gary Olson. "In the early stages, with new or experimental software, there's a breaking-in period and you could lose an entire collaboration if you don't provide good support."

When BIRN first set about creating tools to standardize and calibrate magnetic resonance imaging (MRI) machines at multiple institutions, they solved the local support problem with "BIRN-in-a-box." They shipped an entire integrated hardware system, preloaded and preconfigured with BIRN software, to each of the institutions involved. This minimized the amount of local technical development required.

These days, it's BIRN-on-a-disk, which uses virtualization software—with the same goal. That shift to a much simpler technological model is not insignificant. Indeed, Cummings suggests, in the United States these days, the technology needed for collaboration is already robust. "Most scientists are pretty savvy technology-wise. They can use wikis, email, video and tele-conferencing, instant messaging," he says. "In this day and age, it's hard to see that as an issue." But, he notes, international collaborations might present a different picture if issues around broadband and computer power lead to a tilted playing field.

> "In the early stages, with new or experimental software, there's a breaking-in period and you could lose an entire collaboration if you don't provide good support," says Olson.

However, even in this country there are signs that people are only partly technically ready—and that the technology is only partly there. When researchers at the University of Illinois, Urbana-Champaign, created a collaborative, online research environment called BioCoRE, they thought others would be drawn to the interface for securely managing a number of research projects, sharing files, and scheduling supercomputer time. But, says **Kirby Vandivort**, senior research programmer on BioCoRE, "A lot of people saw BioCoRE as only being a Web interface to their normal tools." It was very difficult to convince people of the value added over email and SSH, a program that lets you securely access and submit jobs on remote computers. So although 2500 people have registered for BioCoRE, only about 50 to 100 folks use it regularly—and approximately half of those are located at the University of Illinois, Urbana-Champaign.

Perhaps the most interesting thing about BioCoRE is how it's evolving. The BioCoRE interface is now built right into the research group's most popular software—VMD, which is used for molecular dynamics visualization. So if researchers want to chat or share a molecule with a collaborator, or schedule supercomputer time, they can now do it entirely within VMD. "This is the real jewel that we didn't actually anticipate when we started," says Vandivort. And it suggests that technology is still finding new ways to make collaboration easier.

MANAGEMENT MATTERS

In larger collaborations, especially at a distance, having a management plan is key, according to the TORSC. "The more seriously the scientists take that plan, working out exactly who will do what. . . the more likely the success," writes Judy Olson.

The tricky thing, Gary Olson says, is that scientists and funding agencies are not keen on spending money on management. They'd rather see the money go to the science itself. But some organizations see the need to take management seriously. BIRN has had a management plan from the

“It’s easy for the home institution to gobble up the lion’s share of the resources with others getting dribs and drabs. But that’s going to kill the collaboration if people feel they’re not being treated fairly,” Olson says.

STARTING OFF RIGHT

A failed collaboration can sometimes lead to legal disputes and misery. To avoid disaster, start your collaboration off right. You might look for Gary and Judy Olson’s soon-to-be-released book, *Science on the Internet*. Or, if you want more of a short course, consult “Guidelines for Negotiating Scientific Collaboration,” published in *PLoS Biology* in June 2005, or “With All Good Intentions,” published in Nature in April 2008, which includes a “collaborators’ prenup.”

COUNSELLING FOR COLLABORATORS

If things start to go sour, get help. The Olsons and their colleagues at the University of Michigan often provide advice to collaborations that are struggling. In addition, Gary Olson now has a grant to create an online tool to help researchers evaluate their own collaborations using the TORSC. He calls it a “Wizard”—a “Collab-o-matic,” if you will. Answer a series of interview questions about your project, and it will give you automatic feedback. The wizard is currently being developed and tested. To Olson, it’s a way to share what his group has learned and also to collect enormous amounts of additional data. So keep your eye out for it. You might be surprised what it will tell you. And what you can tell the Olsons.

get go. They created an oversight committee that, in turn, commissions a variety of standing and ad hoc committees. And BIRN also has as one of its cores the BIRN Coordinating Center.

Good management helps ensure fairness, Gary Olson says. “It’s easy for the home institution to gobble up the lion’s share of the resources with others getting dribs and drabs. But that’s going to kill the collaboration if people feel they’re not being treated fairly. We’ve seen lots of examples of that.”

Another challenge is getting everyone in a collaboration to treat each other as equals. For example, Ellisman points to various NSF-funded cyberinfrastructure projects, in which tension between computer scientists and biomedical scientists impeded collaboration. The result: The computer scientists built something that was underutilized because biologists hadn’t been fully engaged in development.

Ellisman says avoiding this pitfall was his hardest challenge as director of BIRN. “People want to feel recognized for their contribution,” he says. For example, because BIRN needed to involve computer scientists working at the cutting-edge of their own field, it was critical that the biomedical researchers realize the computer scientists’ contributions were equally important—“so they aren’t computer scientists in the service of biomedicine.”

“We started out talking about ‘if you’re a biologist, hug your computer scientist,’” Ellisman says.

Besides fairness, good management also helps build and maintain a collaborative infrastructure. For example, BIRN set out to create a geographically distributed repository of medical images and make them available for large-scale cooperative studies more or less in real time. This involved inventing new software tools to normalize MRIs across multiple institutions; de-identifying health records; and getting approval from the numerous institutional human subject review boards.

“It’s hard work and you take a lot of lumps,” he says. And the coordinating center for BIRN ends up being viewed as a service provider. “Like if the telephone doesn’t work. It’s not how great it is, but what’s not working today,” he says. “It’s part of what happens when you build something that becomes like a utility.”

MODULARITY MATTERS

According to the TORSC, successful collaborations divide the work so that it can be done without the need for a lot of chit-chat. “We have seen a number of projects fail because tightly coupled work spanned people in different locations,” writes Judy Olson and her TORSC colleagues. “The more modularized the work at different locations, the more likely is success.”

Indeed, distance may not matter so much when a collaborative task is easily divided into distinct tasks. And Cummings says this should provide a glimmer of hope to biomedical computing collaborations located at multiple universities. According to his research with the ITRs, projects that produced computer hardware, software and datasets seemed to suffer less from being located at multiple universities.

“I believe there’s something fundamentally different about tools projects,” Cummings says. “Tool development often can be decomposed or broken down, and programmers and other developers seem to have a shared understanding of how to visualize that process.” By contrast, Cummings says, other scientists might not have a clear and common understanding of what a successful outcome would look like or what the steps would be to get there. “The nature of the work—the ability to modularize it and divide it up,” says Cummings—that’s where biomedical computing researchers might have an edge.

The Protein Structure Initiative (PSI) is one prominent collaborative project that, on the surface at least, seemed to divvy up tasks effectively. For example, the PSI funded four large-scale centers in various locations around the country. Each of them then created its own multi-institution pipeline for crystallizing proteins, analyzing their structure, and maintaining and disseminating the data generated. Each institution had a discrete job.

“You have to have independent tasks,” says **Ian Wilson, PhD**, professor of molecular biology at the Scripps Research Institute and director of the Joint Center for Structural Genomics, one of the four PSI centers. But structuring the pipeline in a modular fashion isn’t quite enough, Wilson says. “It has to be seamless.” Making the work flow smoothly took a lot of hard work over the first five years of the project. Some of that involved building relationships among people who didn’t know each other before—establishing “collaboration readiness” after the fact. “We had meetings every week among the cores to discuss how communication could be better and to determine how to move materials through the pipeline more efficiently.” At this point, Wilson says, “We’ve got a fantastically well-integrated team working on this seamless pipeline.”

> “Tool development often can be decomposed or broken down, and programmers and other developers seem to have a shared understanding of how to visualize that process,” Cummings says.

But just as that happened, a new coordination task emerged: in PSI-2 (the initiative’s second five years), the four centers were tasked with working together on communal goals, Wilson says. And the bioinformaticians across the four centers have to work together as well. “We have to decide as a group what to do for 70 percent of the project,” he says. So modularity only got them so far. Ultimately, all of the centers have to agree on how to achieve the overall mission of the PSI.

THE FUTURE OF COLLABORATION

WILL WEB 2.0 CHANGE HOW WE COLLABORATE?

Some social scientists suggest that the shift toward collaborative science mirrors changes in the social order: As global networking becomes the modus operandi in all realms, including business and social life, so too will it become a natural part of science. If this is true, then the culture vultures who see Web 2.0 as the wave of the future—with its MySpaces, Facebooks, and Wikipedias—would also predict that this social climate will affect the scientific endeavor. An impact on collaborative science seems almost inevitable.

According to Cummings, this idea may take some time to be realized. “I think the social structure of science is very resistant to change,” he says. Web 2.0 tools promoting collaboration won’t overcome a culture in which junior faculty are less rewarded for playing a role in a collaboration than they are for producing strong individual research. “You’re looking at hiring and promotion of junior faculty, tenure committees, how departments are structured in universities and labs, norms for sharing and being open about information, journals and their restrictions on intellectual property,” Cummings says. “All of these factors play a much larger role than people who are optimistic about the technology may realize.”

But Ellisman disagrees: “The world people sit at everyday is a collaborative one,” he says. In fact, the BIRN portal right now is a customizable, chattable, bloggable, build-your-own environment type of space. And the next generation will move toward standardization. Just as browsers such as Firefox, Explorer and Netscape have all started to look alike, with similar menus, he says, “for these collaboration spaces we’re going to see a kind of refinement in what is practical and familiar.”

And he sees a gradual shift toward openness as well—a fundamental change in scientific culture. Colleagues Ellisman’s age (60), may still cling to their data, he says. “But for my youngest children, it’s obvious that in the open, available, electronic age where you can exchange music legally or illegally, everything should be open.” Indeed, he says, “We need to err on the side of open access all the way down to the laboratory notebook.”

Some scientists are already there. People are blogging about the potential for open lab notebooks at

OneBigLab.Blogspot.com. The Web site myExperiment.org is giving it a go by allowing researchers to upload workflows so that other people can benefit from them. And in April 2008, Scientific American reported on the phenomenon—pointing to the early success of OpenWetWare.org, a wiki created by graduate students at the Massachusetts Institute of Technology in 2005. Protocols posted on that site have become useful to many other scientists around the world.

“The world people sit at everyday is a collaborative one,” Ellisman says.

As for scientists using Web 2.0 to find collaborators, Nature ran a story in February 2008 titled “The New Networking Nexus.” It describes the spawning of Web sites geared toward bringing researchers together online to discuss their common interests. Examples include Nature’s own site called Nature Network (network.nature.com), which apparently draws interdisciplinary scientists, and Community of Science (COS.com). But, some say, getting a critical mass of participants limits the usefulness of these and other sites.

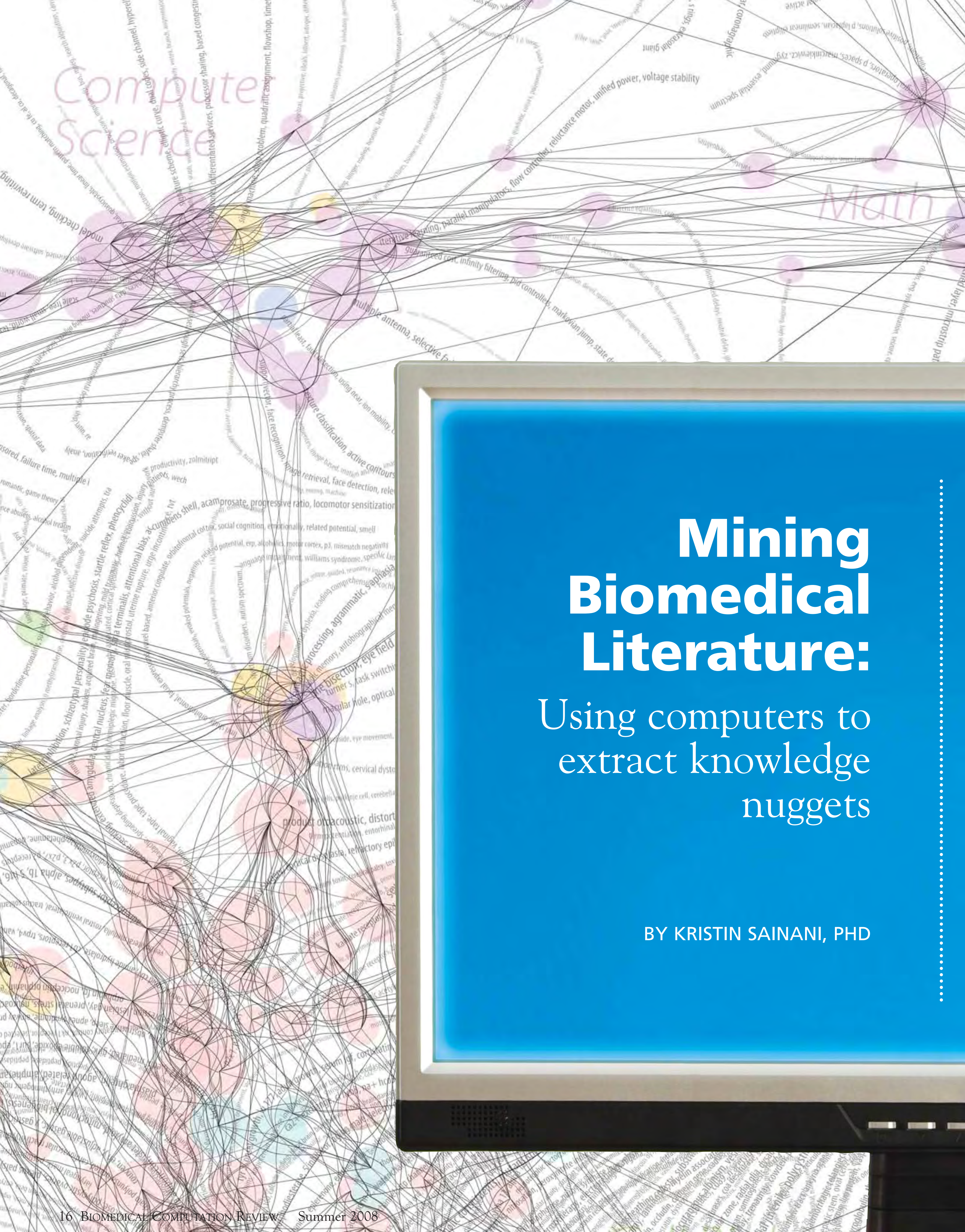
Right now, so many are attempting to develop Web 2.0 tools for science that it’s tough to see where they’ll lead. Tools such as the BioCoRE Web site are created with a specific purpose in mind, but they may end up being used in a different way, or perhaps not being used at all.

Ellisman predicts that Wikipedia is the model of the future. “This is what we expect the world to look like—where all mankind’s knowledge is available to all from anywhere, anytime. That’s what the electronic age we’re poking through right now is going to make possible.”

If he’s right, as data becomes available virtually and search engines become more intelligent, perhaps collaboration among colleagues will become so interwoven with the everyday lives of scientists that it won’t even be called collaboration anymore. But don’t count on that happening anytime soon.

For now, say the social scientists, if you want to succeed, your best bet lies in collaborating with people who work less than thirty meters away and with whom you’ve worked before. Oh, and don’t forget to divvy up the tasks effectively and create a management plan. Though research shows that following these guidelines only gives you a 30 percent chance for a successful collaboration, it is a possibility that can become a reality, even for remote collaborations.

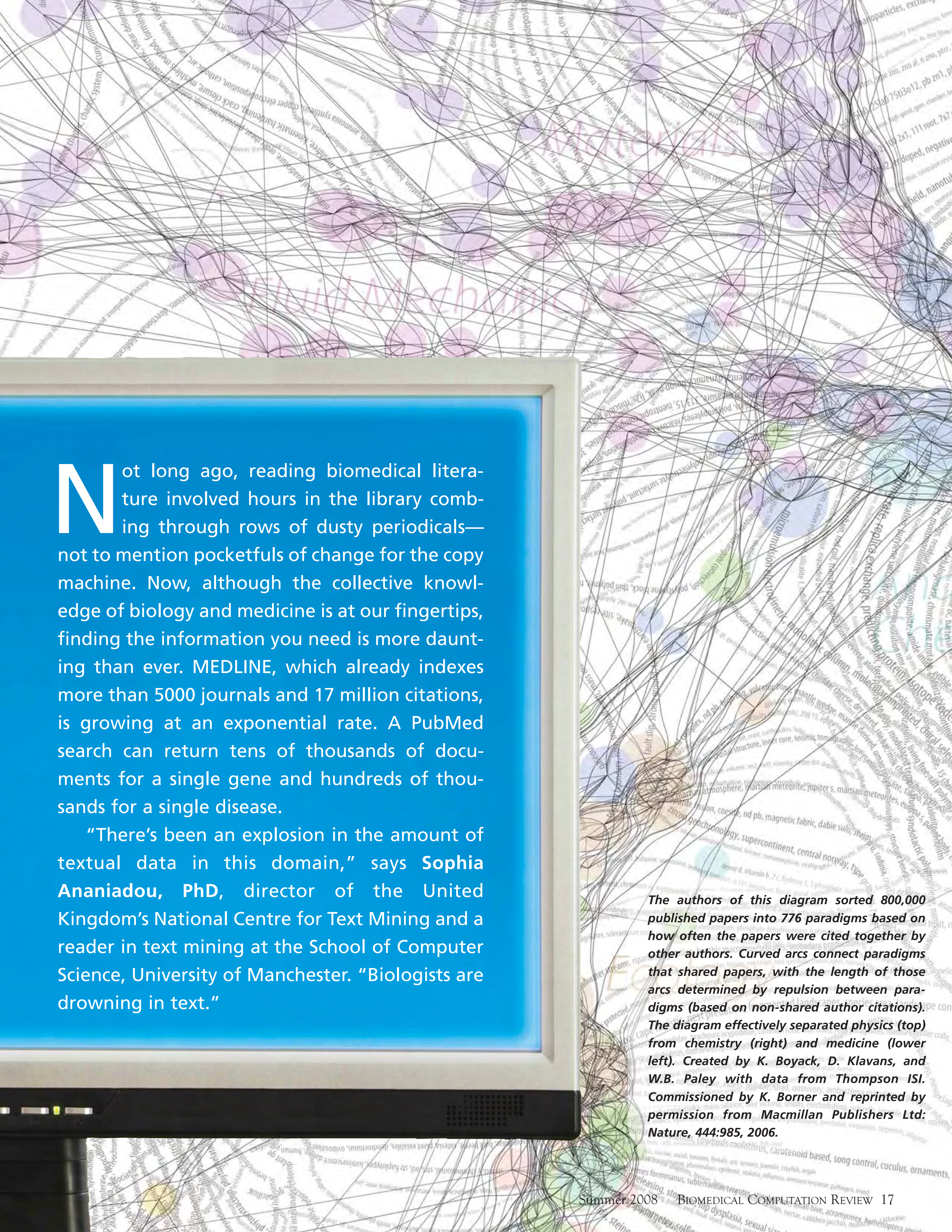
“The biggest reward from running BIRN,” Ellisman says, “has been seeing the acceptance by the biomedical research community of the notion of working across the boundaries of institutions and different domains, and of cooperating to take on larger challenges for the benefit of society and for our understanding of biomedicine and the human predicament.” □



Mining Biomedical Literature:

Using computers to extract knowledge nuggets

BY KRISTIN SAINANI, PHD



Not long ago, reading biomedical literature involved hours in the library combing through rows of dusty periodicals—not to mention pocketfuls of change for the copy machine. Now, although the collective knowledge of biology and medicine is at our fingertips, finding the information you need is more daunting than ever. MEDLINE, which already indexes more than 5000 journals and 17 million citations, is growing at an exponential rate. A PubMed search can return tens of thousands of documents for a single gene and hundreds of thousands for a single disease.

"There's been an explosion in the amount of textual data in this domain," says **Sophia Ananiadou, PhD**, director of the United Kingdom's National Centre for Text Mining and a reader in text mining at the School of Computer Science, University of Manchester. "Biologists are drowning in text."

The authors of this diagram sorted 800,000 published papers into 776 paradigms based on how often the papers were cited together by other authors. Curved arcs connect paradigms that shared papers, with the length of those arcs determined by repulsion between paradigms (based on non-shared author citations). The diagram effectively separated physics (top) from chemistry (right) and medicine (lower left). Created by K. Boyack, D. Klavans, and W.B. Paley with data from Thompson ISI. Commissioned by K. Borner and reprinted by permission from Macmillan Publishers Ltd: Nature, 444:985, 2006.

Electronic access offers the promise that computers might rapidly process and integrate this wealth of information. But the information is recorded in natural language and pictures, which are hard for computers to make sense of. General-purpose text mining tools make a stab at it. But despite substantial progress during the past half century, they are far from giving computers the ability to "read" and understand language in any human sense. Plus, tools developed for general English don't work well when applied to papers containing bioscience jargon.

Fortunately, computational linguists and computer scientists are teaming up with biologists and physicians to develop text-mining tools for biomedicine. "There's been a huge expansion of the field in the past six or seven years," Ananiadou says, including a flurry of papers, competitions, and conference sessions.

Researchers have developed a range of approaches. Some rely on minimal language processing, such as statistical algorithms that look at word counts. Others, by contrast, dig deeper to discern basic language structure and meaning (such as identifying noun phrases or genes) or even reveal the complete grammatical structure of millions of sentences. The latter approach is the most sophisticated and (if perfected) promises to deliver the most precise and comprehensive information, but lower level approaches can deliver a big payoff with much less complexity. Besides mining text, other researchers are working on an arguably more difficult problem for a computer—mining images and diagrams.

The potential applications are as wide-ranging as the biomedical literature itself. Researchers are not simply retrieving and repackaging what is already known, but are also deriving new knowledge by discovering connections that were previously unnoticed. Systems can already generate novel hypotheses by connecting missing links in the literature; predict unknown features of genes and proteins; help researchers make sense of microarray data; extract information to fill biological databases; build large networks of protein, gene, molecule, and disease interactions; evaluate the literature-wide evidence for scientific facts; and trace the evolution of scientific ideas.

Someday, text-mining may even make connections that bridge entire disciplines—from physics to statistics to biology, for example, says **Andrey Rzhetsky, PhD**, professor of medicine and human genetics at the University of Chicago. "You may be able to discover connections between ideas that are far, far away in the knowledge universe," he says.

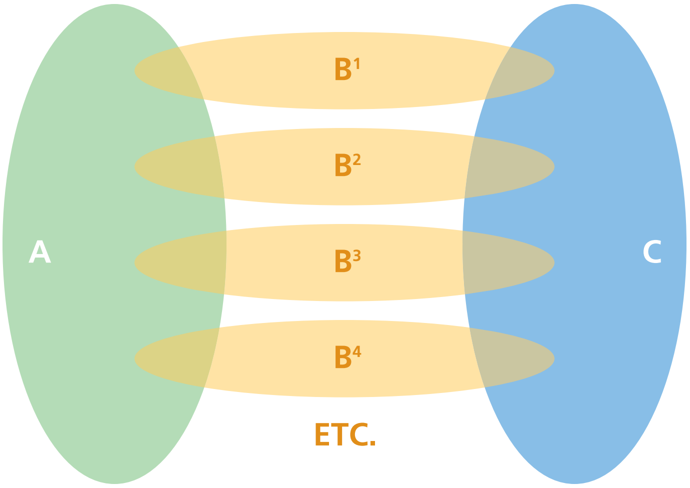
Word Hopping

Connecting the Missing Links

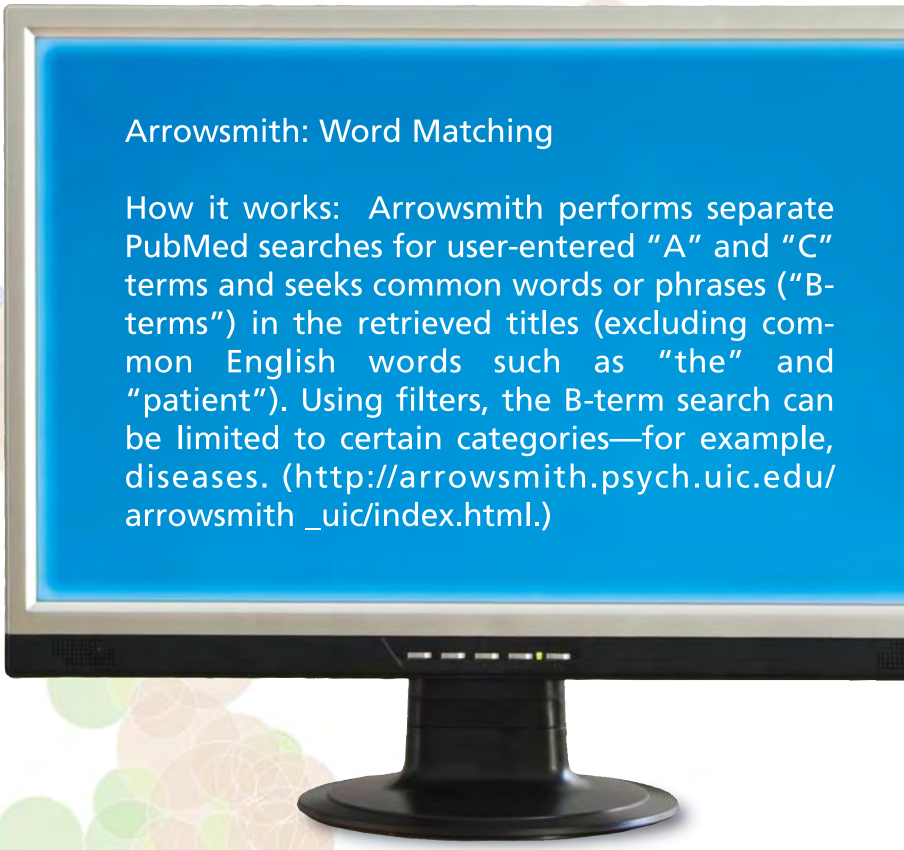
In the 1980s, **Don Swanson, PhD**, now professor emeritus at the University of Chicago, made an early successful attempt to generate novel hypotheses by mining the biomedical literature. He was able to identify indirect connections between therapies and diseases that had never been explicitly linked in the literature. For example, he tied fish oil to Raynaud's disease, and magnesium to migraines. Both treatments were later tested and proven effective.

"One experimental paper may report explicitly that A influences B; another paper, published in some other journal at some other time, may report that B influences C. The inference 'A influences C' will represent an implicit assertion that may be novel, nontrivial and worthy of investigation," explains **Neil R. Smalheiser, MD, PhD**, assistant professor in psychiatry at the University of Illinois at Chicago. He teamed up with Swanson in the 1990s to automate this strategy, creating an online tool called Arrowsmith, which has since been updated and expanded.

Arrowsmith has about 1200 unique users per month. And even though it only parses the titles of papers, Smalheiser says it has already helped researchers formulate new experiments. Among the documented successes, **John Goudreau, DO, PhD**, an associate professor of neurology and toxicology/pharmacology at Michigan State University used Arrowsmith to link Parkinson's disease and Viagra (which had never been studied together): Viagra increases cyclic GMP levels in cells, and cyclic GMP are neuroprotective in several model systems for Parkinson's. He subsequently received a grant from Pfizer to study the association.



***Discovering New Connections:** Arrowsmith finds links (B) between previously unlinked literatures (A and C). For example, a search for "Viagra" (A) and "Parkinson's disease" (C) returned the common phrase "cyclic GMP" (B)—a potential connection between a disease and drug that had never been studied together before. Credit: Vetle I. Torvik and Neil R. Smalheiser, University of Illinois at Chicago*



Arrowsmith: Word Matching

How it works: Arrowsmith performs separate PubMed searches for user-entered "A" and "C" terms and seeks common words or phrases ("B-terms") in the retrieved titles (excluding common English words such as "the" and "patient"). Using filters, the B-term search can be limited to certain categories—for example, diseases. (http://arrowsmith.psych.uic.edu/arrowsmith _uic/index.html.)

Word Profiles

Classifying Genes and Proteins

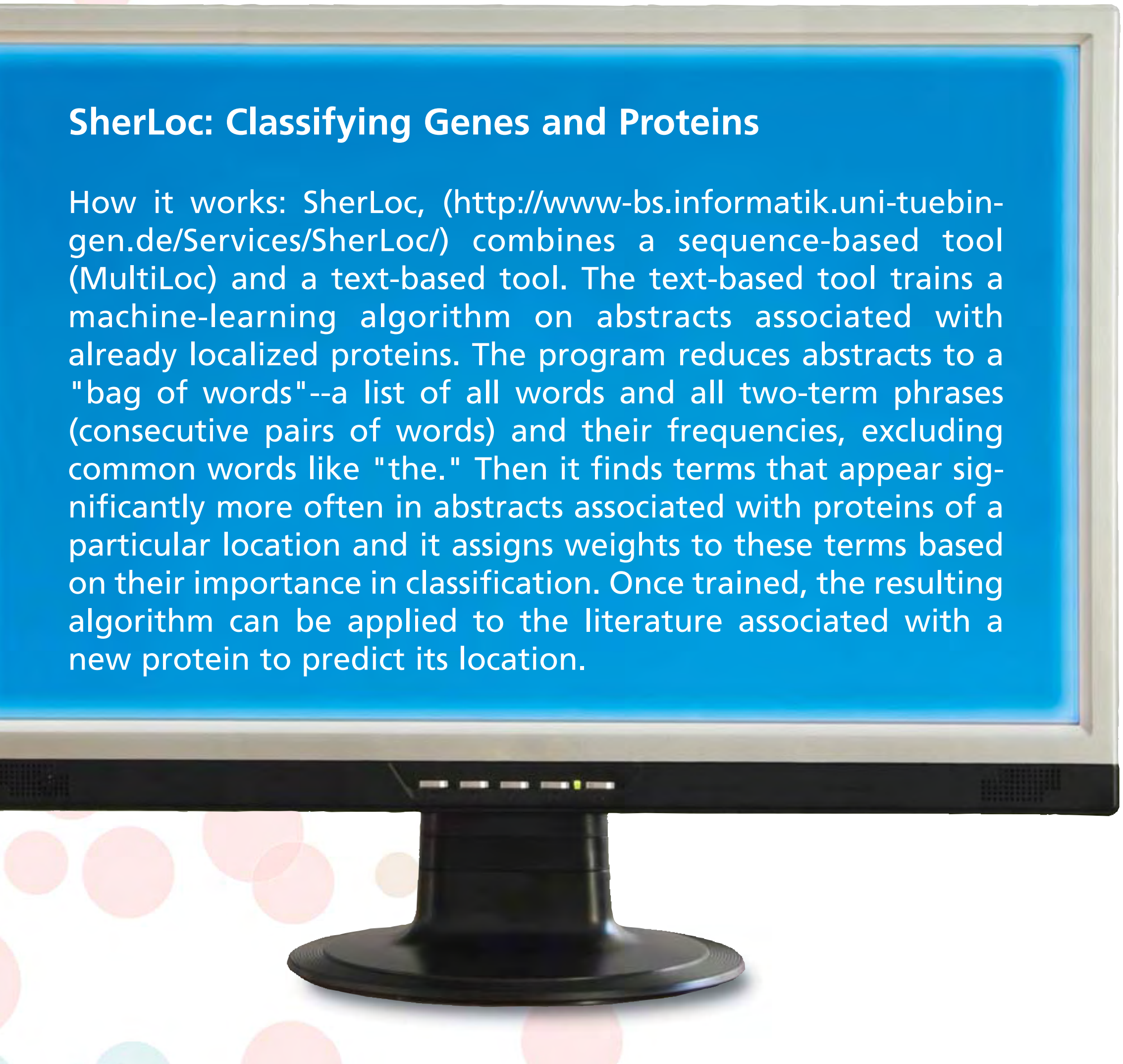
Anyone who has browsed books at Amazon.com has seen some basic text mining in action; for example, books are tagged with "Statistically Improbable Phrases"—terms that occur significantly more frequently in a particular book relative to other books—to give customers one snapshot of a book's essence. A similar strategy can be applied to the medical literature to capture the essence of a gene or protein.

For example, a team led by **Hagit Shatkay, PhD**, an associate professor in the School of Computing at Queen's University in Ontario created a text-based tool to predict where proteins localize in the cell. The idea is that the biomedical literature available for a protein can give clues about its cellular location even before that protein has been localized experimentally. For example, because mitochondrial proteins and nucleus proteins play different roles in the cell, research publications describe them using different terms, Shatkay says. Their program finds terms that occur significantly more frequently in abstracts associated with proteins of a particular location compared with other locations—for example 'bind', 'dna', 'control', 'histone', and 'transcript' for nucleus proteins.

They combined this text-based tool with MultiLoc, a tool that predicts protein localization based on sequence data, which was created at the University of Tübingen in Germany (by a team of scientists led by **Oliver Kohlbacher, PhD**, professor for simulation of biological systems). "MultiLoc, as far as I know, was the most extensive and accurate system at the point where we joined forces," Shatkay says. "The question was could we use text to make it even better?"

Indeed, the integrated tool, SherLoc, gave significantly better predictions of protein localization than MultiLoc alone. Across all organelles, average accuracy for MultiLoc was 74.6 percent and for SherLoc was 85.1 percent (as estimated by cross-validation). "We show that by using text, you can really get an improvement," Shatkay says.

A similar text-based strategy can also be applied to help researchers interpret microarray data, Shatkay says. When a biologist identifies a cluster of co-expressed genes, she can then predict whether they share biological function based on the similarity of their literature profiles. "The advantage of doing it in document space is it gives you some idea of semantics," Shatkay says. If genes cluster based on shared function, the resulting word profile will betray the function (for example, with informative terms such as "fatty acid metabolism"). When clusters form for reasons besides function—such as shared experimental methods—this will be similarly transparent. It's been almost a decade since she and others—including **Steven Edwards, PhD, Mark Boguski, MD, PhD,** and **John Wilbur, MD, PhD**, then all at National Center for Biotechnology Information



SherLoc: Classifying Genes and Proteins

How it works: SherLoc, (http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/) combines a sequence-based tool (MultiLoc) and a text-based tool. The text-based tool trains a machine-learning algorithm on abstracts associated with already localized proteins. The program reduces abstracts to a "bag of words"--a list of all words and all two-term phrases (consecutive pairs of words) and their frequencies, excluding common words like "the." Then it finds terms that appear significantly more often in abstracts associated with proteins of a particular location and it assigns weights to these terms based on their importance in classification. Once trained, the resulting algorithm can be applied to the literature associated with a new protein to predict its location.

(NCBI)—pioneered this strategy, and other scientists are now rediscovering it, she says.

Shatkay's approach involves little natural language processing; it doesn't identify 'this is a gene' or 'this is a noun', for example. But the simplicity is what makes the work elegant. "There are some text-related problems that are relatively easy to solve," Shatkay says. "And the question is, if we solve these problems, can we get anything out of it or do we need to solve the really hard problems before we can get any leverage from text?"

"It won't be as clean, it won't be as nice as natural language processing, but it's really readily available. It's low-hanging fruit," she says.

"There are some text-related problems that are relatively easy to solve," says Hagit Shatkay. "And the question is, if we solve these problems, can we get anything out of it or do we need to solve the really hard problems before we can get any leverage from text?"

Toward Language

Teaching Computers To Read Biology

While statistical approaches yield a big payoff for less effort, researchers in natural language processing are after the holy grail of text mining—getting computers to understand language in some way.

If you just use machine learning and count bags of words while ignoring linguistic structure and meaning, "there's stuff that's just going to stay out of reach," says **Kevin Cohen**, lead artificial intelligence engineer at The MITRE Corporation and biomedical text mining group lead at the University of Colorado School of Medicine.

But tackling natural language is enormously difficult. "There's this sense, this assumption, that it should be easy. You can talk and understand things and read things really easily. But of course your whole brain is designed for that," says **Alex Morgan, MS**, a doctoral student in biomedical informatics at Stanford University. "And you think that things like analytical chemistry and scheduling of flights are really complicated problems, but those are trivial computer problems [compared with natural language processing]."

To incorporate language, text-mining researchers use extensive lexical resources (including word lists, thesauri, and ontologies) to look up word variants and meanings; manually created rules about grammar and language; and machine-learning algorithms trained on collections of text marked up with linguistic information (annotated corpora).

In the world of news and journalism, considerable progress has been made on two key language-based tasks: identifying simple entities such as places, organizations, and people; and extracting simple facts such as "Company A took over Company B." Researchers have achieved near human proficiency on the first task, evaluated with F-measures—a quantity that combines precision (getting it right) and recall (not missing anything)—above 95 percent; and reasonable performance on the second task, with F-measures of 70 to 80 percent. But when off-the-shelf systems were applied to biology, they did poorly. Having been trained on text from the news world, such as The Wall Street Journal corpus, they were ill-equipped to tackle biomedical journal articles written by scientists and containing considerable jargon and nonstandard grammar.

Fortunately, in the past decade, several key events have advanced natural language processing in the biomedical domain. First, researchers in Japan—led by **Junichi Tsujii, PhD**, professor of computer science at the University of Tokyo and professor of text mining at the University of Manchester in the United Kingdom—created the GENIA corpus, a collection of PubMed abstracts annotated with both linguistic and biological information. Corpus-based techniques had revolutionized natural language processing, Tsujii says. "I thought I should apply a similar approach to biotext mining."

"Then we made that corpus available to all the researchers in the world," he says. "And I think that contributed quite a lot to the progress of bio-text mining."

Another driving force was the creation of a series of challenge evaluations (competitions) for text mining in biology called BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology), which started in 2003 and is run by The MITRE Corporation and the Spanish National Cancer Research Center (CNIO). Challenge evaluations can help drive a field forward by creating resources, building a community of researchers, and providing standards for assessment.

"At the time, I found myself asking: What's the state of the art with respect to text mining for biology? And, if we can do 90-plus percent accuracy on newswire, why don't we get that performance in biology?" says **Lynette Hirschman, PhD**, director of biomedical informatics at The MITRE Corporation. There was also a need for a standard way to assess text-mining tools and a need to assess them on datasets other than the ones that were used to train them. Results on researchers' private datasets were all over the map, says Hirschman. The BioCreAtIvE competition was intended to fix that problem.

As described below, BioCreAtIvE has also addressed key challenges in bio-text mining—including promoting the development of tools to find gene and protein mentions in text and extracting basic facts, such as protein-protein interactions.

"There's this sense, this assumption, that it should be easy [for computers to read natural language]. You can talk and understand things and read things really easily. But of course your whole brain is designed for that," says Alex Morgan.

The Name Game

Tagging Genes and Proteins

One of the most basic tasks in natural language processing is to recognize important entities in running text. In biology, this means identifying genes, gene products, diseases, drugs, and cells and linking them to a unique identifier (such as an EntrezGene or SwissProt ID). If this foundational task is done poorly, the accuracy of higher-level tasks suffers.

"It's a very pragmatic problem, but it's very hard in the biomedical domain. It's surprisingly much easier in economics or business or news, because categories are better defined and less overlapping," Rzhetsky says. "But here it's essentially a mess."

Gene and protein names present a particular challenge. Historically, scientists have used whimsical names that are not readily distinguishable as genes, for example: cheap date, heartless pinhead, and Indy (short for "I'm not dead yet"). A gene or protein may also have multiple name variants, for example, S-receptor kinase with and without the hyphen; or nuclear factor kappa B and NFKB. Further, it may be hard to distinguish between a gene and its gene product; for example, 'p53' could refer to a gene, protein, or mRNA. Gene and protein names may also be shared across species. "Those things never happen in the newswire domain. Bill Clinton is always Bill Clinton," Tsujii says.

The bio-text mining community has focused considerable attention on this problem in the past five years. Several named entity recognition tools for biology are publicly available, such as those provided by the United Kingdom's National Centre for Text Mining (http://www.nactem.ac.uk/), a centre created to provide text-mining tools and services for biologists. Existing tools draw on dictionary lookup (matching strings in text with lists of names); manually constructed rules, such as 'any word ending in ase is a protein' or 'any phrase containing the word receptor is a protein'; and machine-learning techniques.

The top systems in BioCreAtIvE—which all include machine-learning components—achieve F-measures of 80 to 90 percent for finding gene mentions and normalizing these genes to a unique ID, which represents the state of the art for this task.

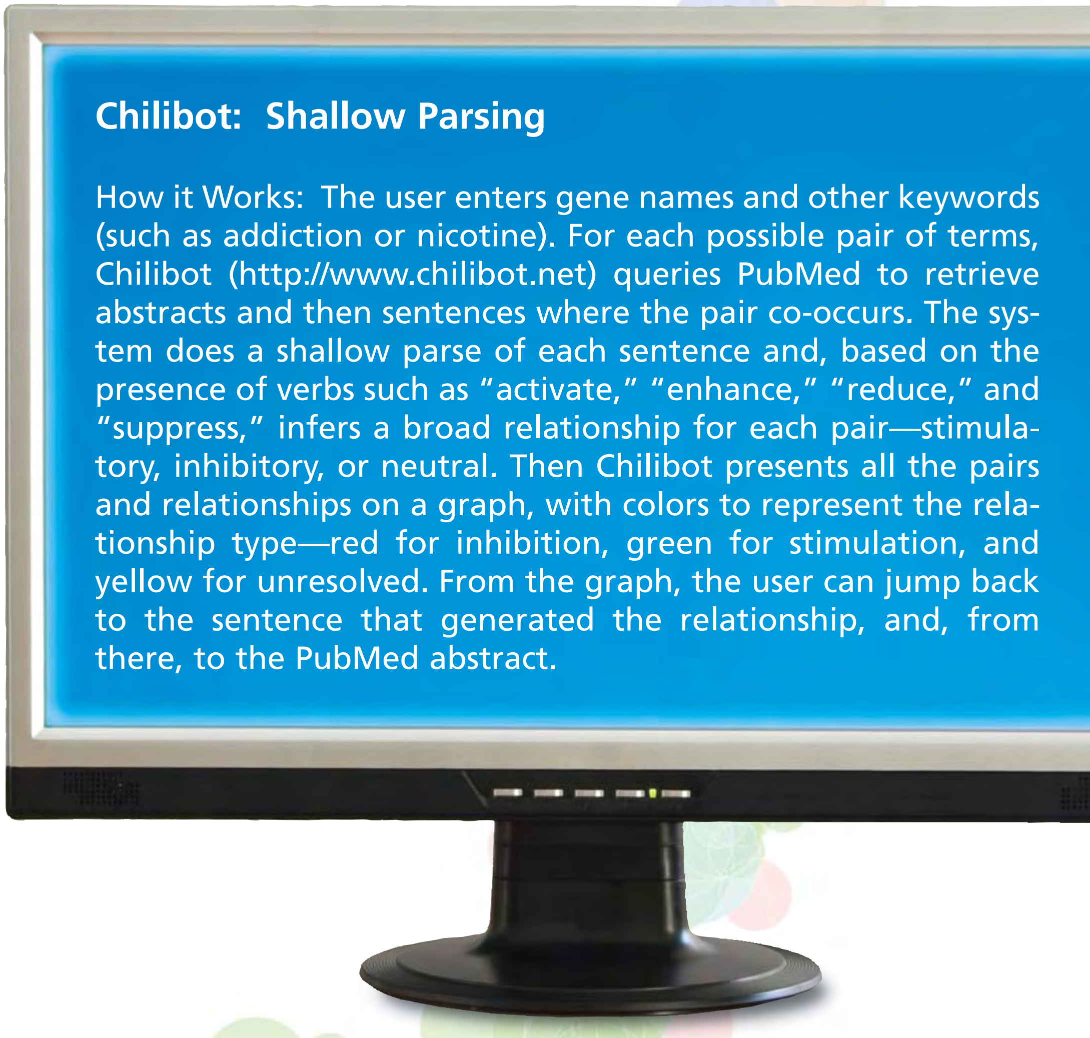
What's promising is that the best systems in BioCreAtIvE 2006-2007 surpassed those in 2004, Hirschman says, and progress should continue.

Sentence Slicing and Dicing

Mining for Relationships

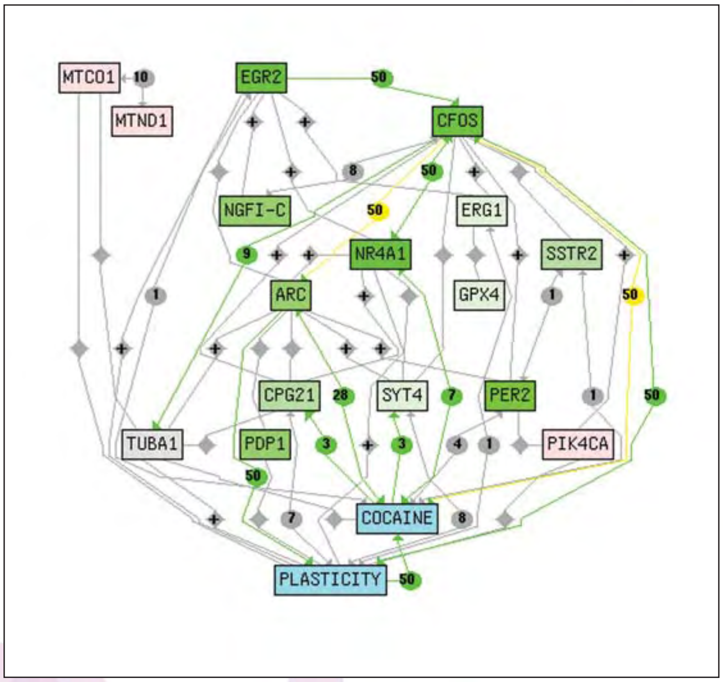
The next step is to extract simple facts, such as protein-protein interactions, gene-disease relationships, and drug-gene relationships. These facts can be used to fill biological databases or to reconstruct biological pathways.

Fact extraction systems in biology use various degrees of "parsing"—teasing out a sentence's grammatical structure. Early systems used no parsing, but simply inferred interaction when two proteins (or other entities) appeared in the same sentence (co-occurrence). Later systems used shallow parsing—identifying noun and verb phrases—to find telltale patterns such as two noun phrases around the verbs "phosphorylate," "bind," or "activate." The latest trend is



Chilibot: Shallow Parsing

How it Works: The user enters gene names and other keywords (such as addiction or nicotine). For each possible pair of terms, Chilibot (http://www.chilibot.net) queries PubMed to retrieve abstracts and then sentences where the pair co-occurs. The system does a shallow parse of each sentence and, based on the presence of verbs such as "activate," "enhance," "reduce," and "suppress," infers a broad relationship for each pair—stimulatory, inhibitory, or neutral. Then Chilibot presents all the pairs and relationships on a graph, with colors to represent the relationship type—red for inhibition, green for stimulation, and yellow for unresolved. From the graph, the user can jump back to the sentence that generated the relationship, and, from there, to the PubMed abstract.



Building Networks: The Chilibot program mines PubMed abstracts for broad relationships (stimulatory, inhibitory, or neutral) between genes, proteins, drugs, and biological processes and presents them graphically. Here, Chilibot summarizes how a group of genes relate to each other and to the biological concepts of plasticity and cocaine. Credit: Hao Chen, University of Tennessee Health Science Center

to use deep parsing—specifying the full grammatical structure of a sentence—to unravel nested and complex relationships and deal with more complex grammar (such as passive voice).

A widely used fact-extraction system that employs shallow parsing is Chilibot, short for "chip literature robot." The program constructs relationship networks among biological concepts, genes, proteins, and drugs, and presents them in graphical form.

"You provide a list of terms and then you retrieve a graph of highly summarized relationships between the terms," says **Hao Chen, PhD**, assistant professor of pharmacology at the University of Tennessee Health Science Center. Chen, a neurobiologist, wrote the program to help him interpret microarray data. It will show how a list of co-expressed genes connects with each other and with a biological process, such as addiction.

"Anything that can synthesize all the literature is presumably better than something that only looks at a little bit," Alex Morgan says.

Chilibot has been used in the design, interpretation, and validation phases of experiments, Chen says. "There are many programs with a similar function. But Chilibot probably is one of the most user-accessible interfaces of this technology on the web," he says.

Shallow parsing has limitations, though. "Basically, you get what you pay for," says Ananiadou. Deeper parsing delivers more precision and handles complex, nested chains of interactions. For example, the sentence "Phosphorylated Cbl coprecipitated with CrkL, which was constitutively associated with C3G" involves several nested relationships that can only be correctly mapped out with deep parsing. "If you want to work on systems biology, with pathways, you need to go to a much deeper level," Ananiadou says. "So this is where the community is moving now."

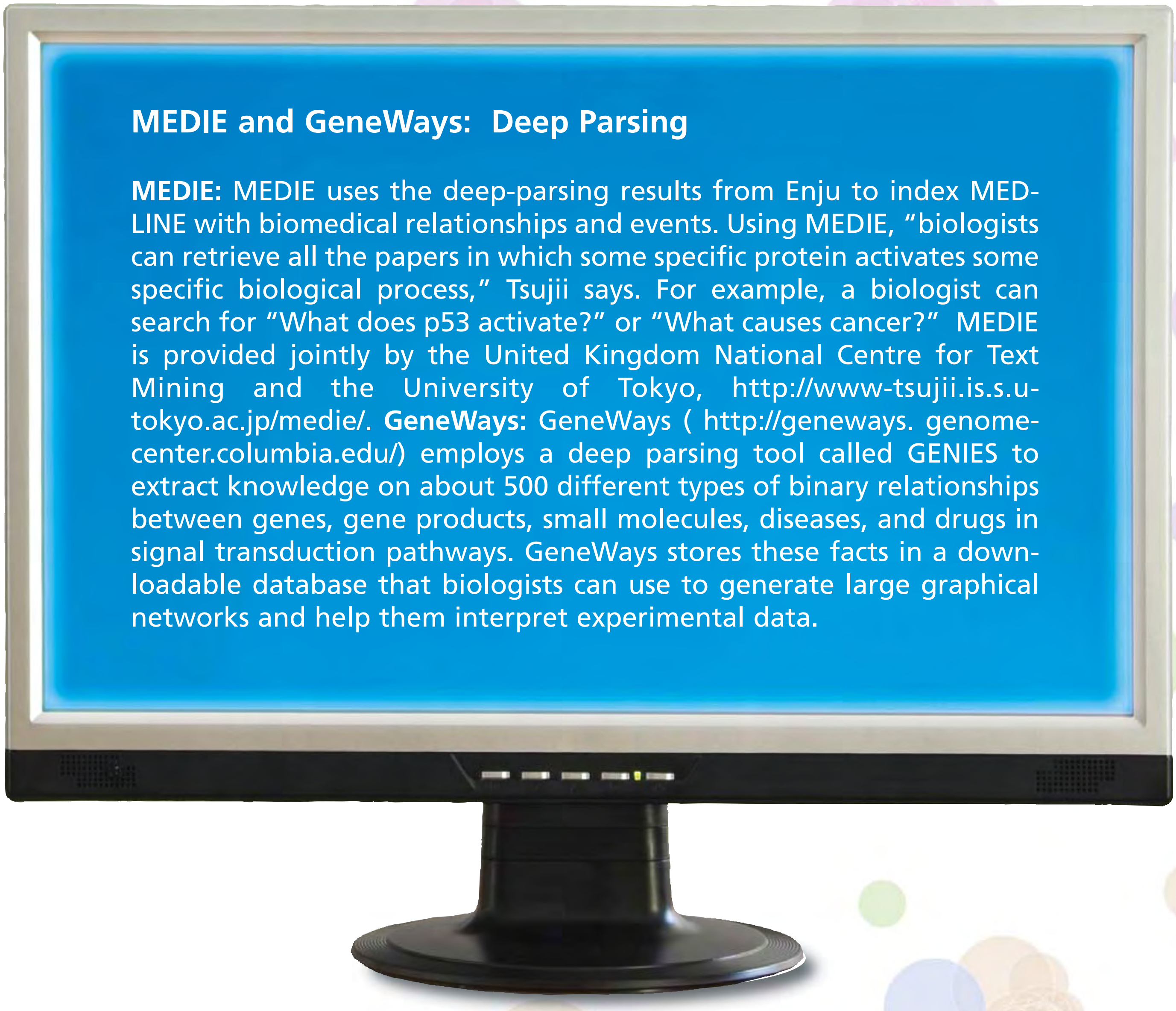
Not only are researchers trying to achieve depth, they are also trying to achieve breadth. Some groups have actually parsed the whole of MEDLINE and beyond. "Anything that can synthesize all the literature is presumably better than something that only looks at a little bit," Morgan says.

For example, Tsujii's lab developed a deep parsing tool called Enju (http://www-tsujii.is.s.u-tokyo.

Open access publishers—such as PubMed Central and PLoS—have unlocked a critical door for bio-text miners by providing full-text articles in a computer-friendly format.

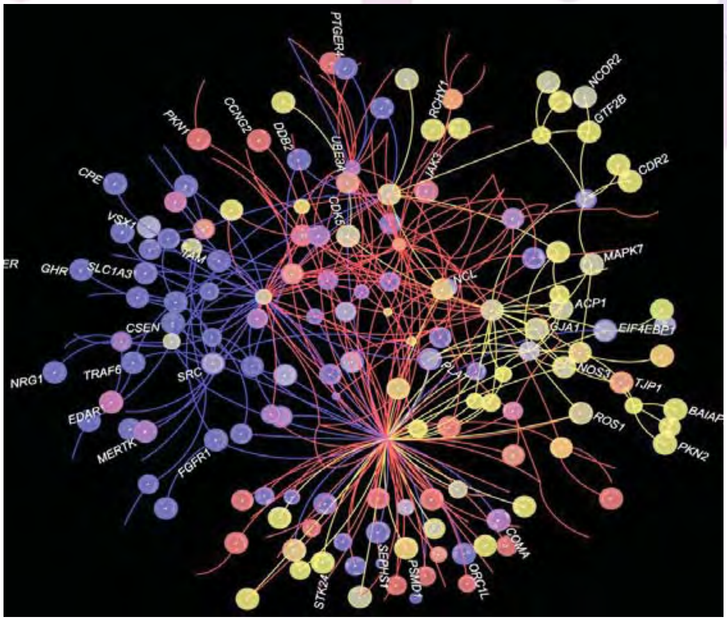
ac.jp/enju/index.html), named after a Chinese tree of wisdom (since it generates syntax trees). Enju is a cutting-edge system, one of the few text-mining programs in biology that does full parsing, Tsujii says. Tsujii's team used Enju to parse all 70 million sentences in MEDLINE in about eight days (using a 350 PC cluster). From there, they extracted all the biomedical events (such as protein-protein interactions) reported in MEDLINE; these results form the basis of an intelligent search tool called MEDIE. "Our next goal is to map all the events reported in MEDLINE to some kind of complex network," Tsujii says.

Another cutting-edge program that employs literature-wide deep parsing is GeneWays (http://geneways.genomecenter.columbia.edu/), developed at Columbia University. GeneWays extracts knowledge on biological relationships in signal transduction pathways and puts these facts into a database that biologists can download. The latest run (which took about 3 months) parsed about one-third of a million full-text articles from 100 peer-reviewed journals as well as all of PubMed, and generated about 8 million redundant



MEDIE and GeneWays: Deep Parsing

MEDIE: MEDIE uses the deep-parsing results from Enju to index MEDLINE with biomedical relationships and events. Using MEDIE, "biologists can retrieve all the papers in which some specific protein activates some specific biological process," Tsujii says. For example, a biologist can search for "What does p53 activate?" or "What causes cancer?" MEDIE is provided jointly by the United Kingdom National Centre for Text Mining and the University of Tokyo, http://www-tsujii.is.s.u-tokyo.ac.jp/medie/. **GeneWays:** GeneWays (http://geneways. genomecenter.columbia.edu/) employs a deep parsing tool called GENIES to extract knowledge on about 500 different types of binary relationships between genes, gene products, small molecules, diseases, and drugs in signal transduction pathways. GeneWays stores these facts in a downloadable database that biologists can use to generate large graphical networks and help them interpret experimental data.



Pulling Out Pathways: The GeneWays program parses the biomedical literature and returns millions of published relationships between genes, gene products, small molecules, diseases, and drugs. Here, researchers mapped the relationships between genes believed to be involved in autism (blue), bipolar disorder (yellow), and schizophrenia (red) to look for genetic overlaps between the three diseases. Credit: Ivan Iossifov, Columbia University, and Andrey Rzhetsky, University of Chicago

and 4 million unique facts, says Rzhetsky (who helped develop GeneWays). In addition to helping researchers interpret experimental data, Geneways can be used to trace the evolution of ideas in the scientific literature or help derive consensus from conflicting statements in the literature. For example, Rzhetsky's team is working on an algorithm that evaluates the weight of evidence supporting or contradicting a particular fact and generates a probability that the fact is true. "You can try to reconstruct truth," Rzhetsky says.

The performance of state-of-the-art fact extraction systems in biology is unknown, but—on a fact-by-fact basis—it may be low. The top systems on a protein-protein interaction task in BioCreAtIvE 2006-2007 achieved F-measures of only about 35%. Though fact extraction remains largely a research problem, tools in use today are benefiting users. These systems exploit redundancy (looking at multiple mentions of a fact) to increase recall and accuracy.

If you combine information extracted from 50,000 paragraphs, "you're going to get the right answer," Morgan says. "Eventually you're going to have seen that fact so many times that it must be true and all the ones that are wrong disappear, because they're random instances that don't happen that often."

Full Text Ahead

Advancing Biology and Medicine

The bio-text mining community faces several key challenges. To date, tools have focused on mining abstracts, which are more readily available than full-text articles. But the bulk of information, as well as the tables and figures, are contained in full text.

Many full-text articles require a subscription for access; and even when available, they may be in formats that don't work well for text-mining applications.

"Trying to process a PDF document is a nuisance. You can convert it to plain text but it doesn't convert very well," Hirschman says.

Open access publishers—such as PubMed Central and PLoS—have unlocked a critical door for bio-text miners by providing full-text articles in a computer-friendly format. But much of the literature still remains inaccessible.

Another challenge is making tools that are useful to biologists. Systems are typically evaluated as to their recall and accuracy in handling canned problems, but usability to biologists may actually be a more important benchmark.

"We've been pushing for evaluations that will let us quantify the value of a particular system and its performance to a biologist. How much will it help you do your job?" Cohen says. "I'm happy to say that in the last couple years, for the first time, we've actually seen productive research in that area."

Despite such challenges, bio-text mining has advanced considerably in a short amount of time. Rather than scientists tracking down one journal article at a time in the library, computers are now doing the legwork—surveying millions of abstracts and hundreds of thousands of full-text articles at once and returning insights that don't exist in a single article.

"We've made enormous progress," Hirschman says. "We have a very vibrant community of researchers now. The results are getting better. We understand where we are and what resources we need." And if key challenges, such as full-text access and usability, can be met, she and others expect the field to advance rapidly.

Moreover, says Rzhetsky, "I strongly believe that text mining can speed up scientific progress." □

Getting the Picture

Mining Images and Diagrams

While considerable effort has gone into processing biomedical text, much less attention has been paid to processing figures. Yet figures and figure-related text (captions and text referring to figures) make up 50 percent of a typical biomedical paper, says **Robert P. Futrelle, PhD**, associate professor of computer and information science at Northeastern University.

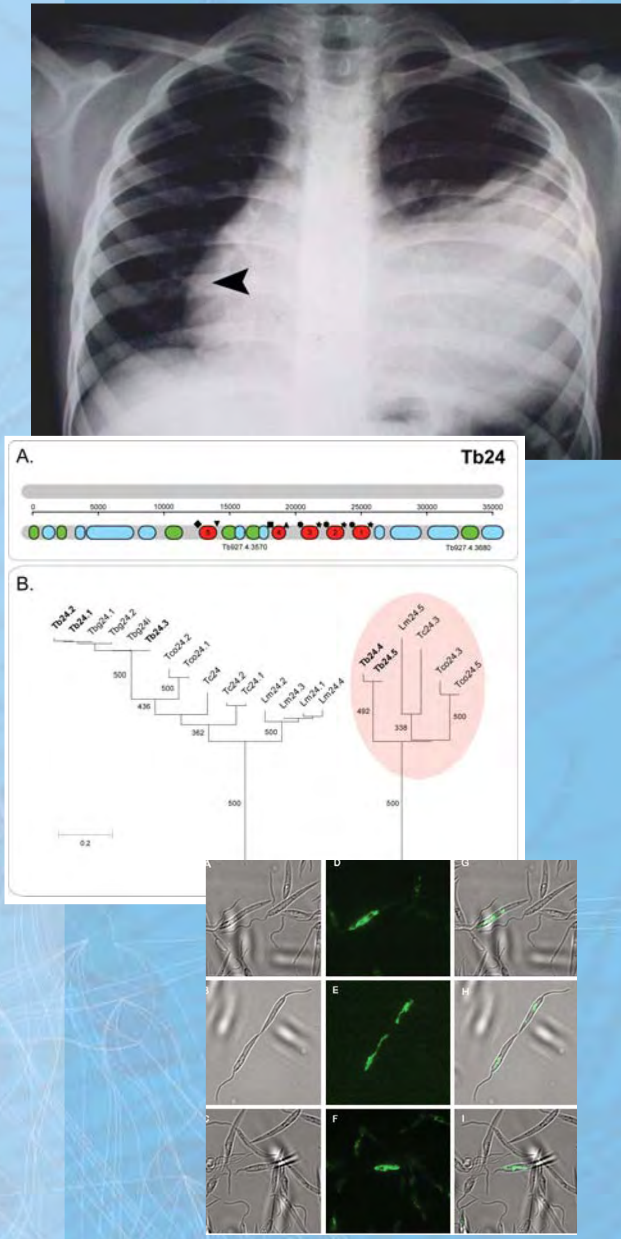
Technology for mining figures is still in its early stages, and most of the work is focused on information retrieval—the ability to search for images and diagrams the way we search for documents. As with text mining, the simplest approach is to use statistical methods in which programs look for patterns of pixels rather than patterns of words. But so far this just allows researchers to classify images in broad terms. For example, in a search of medical images, this technique might separate a chest X-ray from a CT scan. “The technology of image processing is not nearly as advanced as text processing,” says **William Hersh, MD**, professor of medical informatics and clinical epidemiology at Oregon Health & Science University.

Besides medical images, other researchers are working on classifying biological images. For example, a team led by **Robert Murphy, PhD**, professor of biological sciences, biomedical engineering, and machine learning at Carnegie Mellon University, developed SLIF (Subcellular Location Image Finder, http://slif.cbi.cmu.edu), a tool that divides multi-part figures into individual panels and picks out fluorescence microscope images (using a machine-learning classifier). Beyond classification, the tool also extracts facts about protein subcellular localization from image features and caption text. SLIF was used to automatically extract fluorescence microscope images from 15,000 PNAS papers and to store them in a searchable database indexed (where possible) by protein, cell type, and subcellular location.

Futrelle is trying to do even deeper processing—akin to parsing sentences—to extract meaning from diagrams (line drawings and graphs). In diagrams, the lowest level items are not words, but individual lines that have essentially no meaning on their own, he says. “If you just had the lines in a bag and pulled them out it wouldn’t mean anything; but if you put them in place, all of a sudden, ‘bingo,’ you have something,” he says.

Rather than look for nouns, verbs, and prepositions, his team looks at the lengths, positions, and connectivity of lines to detect standard pictorial expressions, such as plus signs, arrows, and error bars. “So, we have parsed diagrams. We have taken data graphs and pulled out everything—all the little tick marks and the scale lines and the data points,” he says. His lab is now redeveloping the approach in a newer programming language.

In the future, Futrelle says he hopes to build tools that perform intelligent searching for particular types of diagrams (such as a bar graph about a specific topic) and that automatically add metadata—tags that identify: “this is a gene diagram” or “this is a bar graph”—to figures in the literature. Beyond information retrieval, Futrelle’s work could also form the basis of systems that actually mine figures for new knowledge, similar to current text-mining systems.



Scientists are making progress mining information from figures such as these. Chest Xray reprinted from Marashi SM, Eghtesadi-Araghi P, Mandegar MH. A large left ventricular pseudoaneurysm in Behçet's disease: a case report. BMC Surg. 2005 Jun 14;5:13. Flourescence Microscope image reprinted from: Zamora-Veyl FB, Kroemer M, Zander D, Clos J. Stage-specific expression of the mitochondrial co-chaperonin of Leishmania donovani, CPN10. Kinetoplastid Biol Dis. 2005 Apr 29;4(1):3. Diagram example reprinted from: Jackson AP. Tandem gene arrays in Trypanosoma brucei: comparative phylogenomic analysis of duplicate sequence variation. BMC Evol Biol. 2007 Apr 4;7:54.

BY JOY KU, PHD

OpenMM: Bringing GPU Acceleration Capabilities to Molecular Dynamics

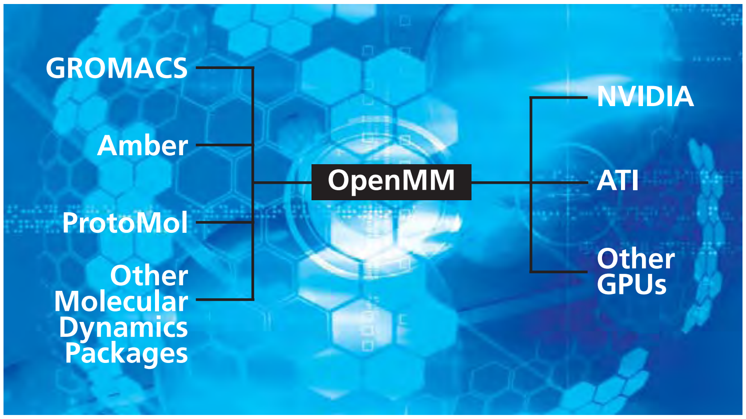
Over the last three years, the lab of **Vijay Pande, PhD**, at Stanford University has optimized their molecular dynamics (MD) algorithms to take advantage of the fast computing that's possible with GPUs, or graphics processing units (see this issue's Under the Hood column for more information about GPUs). Now, through their collaboration with Simbios that capability will be made freely available to the whole community via a library called Open Molecular Mechanics, or OpenMM.

"OpenMM will be a tool that unifies the MD community," says **Russ Altman, MD, PhD**, principal investigator of Simbios and a professor of bioengineering, genetics, medicine, and computer science at Stanford University. "Instead of difficult disparate efforts to recode existing MD packages to enjoy the speedups provided by GPUs, OpenMM will bring GPUs to existing packages and allow researchers to focus on discovery."

There are tens of MD packages available today: GROMACS, NAMD, and Amber to name just a few. Currently, if an applications developer wanted to accelerate their MD software using GPUs, they would have to write multiple versions of their code since each GPU manufacturer uses a different set of commands. OpenMM would provide a common interface, hiding the details of programming the different GPUs.

"The user wouldn't even have to think about any of the GPU craziness," says Pande, associate professor of chemistry and of structural biology at Stanford University and lead of the OpenMM project. "All they would know is that they want to do a force calculation or an energy calculation and they'll just know it's going to be done fast on a variety of hardware."

How fast? "On GPUs, we routinely get speedups by a factor of 100 and in some cases, close to a factor of 1000," says Pande. "Those types of speedups can really change how your work gets done. Things that used to take three years can now get done in a day."



OpenMM makes it easy to use GPUs to speed up different molecular dynamics packages.

Grant Krafft, PhD, Chairman and Chief Science Officer of Acumen Pharmaceuticals, Inc., has benefited directly from faster simulations. His company uses Pande's simulation software to help them design molecules to treat and prevent Alzheimer's.

"With the expanded capabilities of these simulations, we can get a more complete picture of which molecular assemblies prefer to form," Krafft says. "What's really important is that they don't incorporate approximations that many other molecular dynamics calculations have to incorporate, approximations that would lead to errors."

OpenMM makes it possible for other scientists to achieve similar results with their preferred MD code without much more programming and without an expensive supercomputer or a cluster. The only additional hardware that might be needed would be a high-end GPU board, which costs just a few hundred dollars these days and is straightforward to install.

The first release of the OpenMM library is planned for the fall of 2008. The release will include integration of the OpenMM library into the GROMACS MD package.

"Nobody's really coming close to what Vijay's doing in terms of duration of folding and dynamics studies," says Krafft. But with the release of OpenMM, those capabilities could easily become available to all. □

DETAILS

OpenMM is part of Simbios' new protein folding driving biological problem (DBP). Hundreds of the protein folding trajectories generated by the Pande lab are also being made available as part of this DBP. See https://simtk.org/home/foldvillin.

To learn more about OpenMM, visit https://simtk.org/home/openmm. The first open code release of OpenMM is planned for Fall 2008 and will be available for download from this Web site.

Simbios (http://simbios.stanford.edu)
is a National Center for Biomedical Computing
located at Stanford University.

BY JOHN MELONAKOS

Parallel Computing on a Personal Computer

Anyone who has ever waited minutes, hours, or even days for software to complete a biomedical computation will be happy to hear that almost every personal computer is capable of better. Today, most standard PCs, both desktops and laptops, come with a graphics processing unit (GPU) in addition to the central processing unit (CPU). And, thanks to the video gaming market, GPU hardware has advanced at a much faster pace than CPU hardware. In fact, GPUs have advanced so quickly that today they have ten times more computational power than CPUs (see the graph).

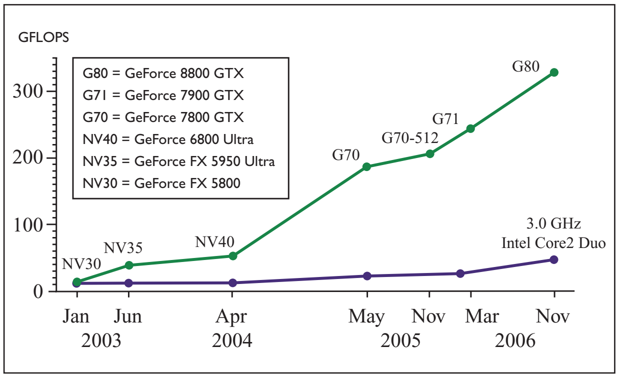
Why are GPUs faster than CPUs for most biomedical computations? A CPU is a serial computing device, processing data sequentially. A GPU is a parallel computing device, processing many chunks of data all at the same time. Since most biomedical computations are parallelizable, GPU computing provides a powerful alternative to traditional CPU computing without the expense of purchasing a room full of clustered computers.

The big bottleneck for GPU computing is writing software specialized for the GPU. Since GPU computing is in its infancy relative to CPU computing, only a small fraction of programmers around the world are familiar with GPU-based programming languages such as CUDA, Brook+, or Ct. GPU software developers must scale a serious learning curve if GPUs are to serve the mainstream.

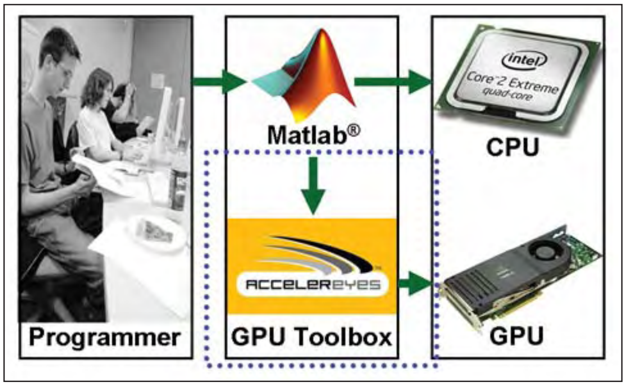
To solve this problem, easy-to-use GPU programming toolboxes are now available, such as the Matlab-based one from AccelerEyes (see the framework). The utility of these tools is to help researchers tap into the benefits of GPU computing.

But GPUs are already having an impact in biomedical computing. Examples include image-guided brain surgery, molecular dynamics simulations, and genomics. Complex algorithms which take hours when run on a CPU can now be used in real-time. And the computing power made available by a GPU on a standard PC now costs hundreds of times less than that of a cluster of PCs having similar computing power.

With these kinds of speed improvements and cost benefits, GPU programming is sure to become mainstream. It's clearly faster than running software on your CPU (especially when that same computer already has the hardware necessary to go faster); and it's clearly cheaper than buying a room full of clustered computers. Now the software world just needs to catch up. □



Over the last few years, the power of GPUs had increased dramatically compared to that of CPUs, as shown in this chart comparing NVidia graphics processors with Intel processors.



AccelerEyes is a new programming tool that allows researchers to use GPUs for Matlab tasks. Courtesy of John Melonakos.

DETAILS

John Melonakos, a PhD student at Georgia Tech, is an active participant in the National Alliance for Medical Image Computing (NA-MIC), one of the National Centers for Biomedical Computing. He joined with Tauseef ur Rehman, Gallagher Pryor, and James Malcolm to start AccelerEyes LLC, which is developing technologies that enable CPU-based code to run on GPUs. The AccelerEyes Jacket Product, connecting Matlab to the GPU, is available by visiting www.accelereyes.com. For more information or to inquire about joining the AccelerEyes team, please send an email to: john.melonakos@accelereyes.com.

DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation
REVIEW
Published by Simbios, a National Center for Biomedical Computing

Biomedical Computation Review
Simbios A National Center for Biomedical Computing
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

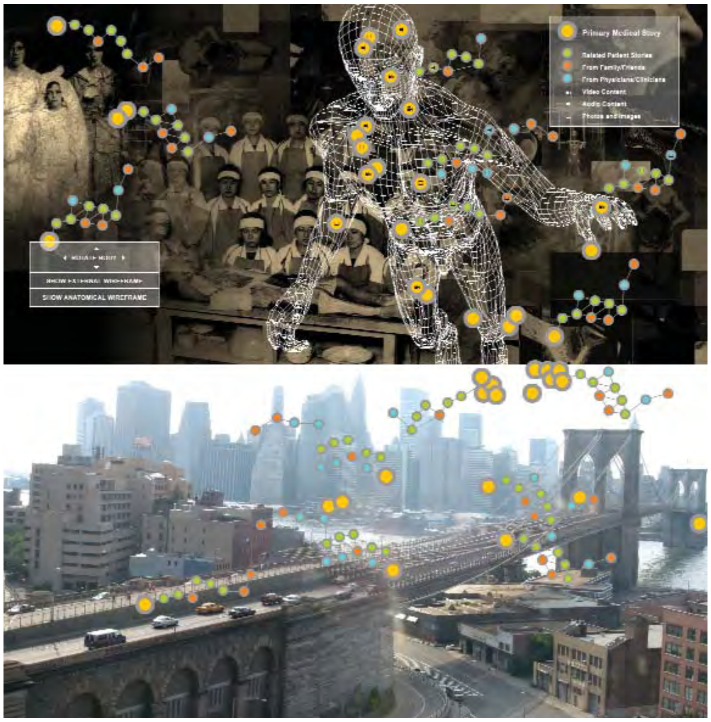
An Avatar of Human Health

The mesh body of a human form floats over the Brooklyn Bridge. Dots of color, embedded with video testimonials, share the collective health problems of 9/11 survivors. In this incarnation, PhineasMap is pure art. "I was attempting to visualize a single, public body that we can all connect to in some manner," says **Virgil Wong**, "as a common point to empathize with the pain of other people."

PhineasMap is now evolving into a health encyclopedia, an avatar for individual human health, and a representation of the collective health of people in a particular location.

"We're currently using this model as a way to archive and access health information," says Wong, the Web Center director at NewYork-Presbyterian Hospital and Weill Cornell Medical College.

Initially, the PhineasMap body will be used as an access point for general health information. He is connecting it to the NewYork-Presbyterian and Weill Cornell online health encyclopedia, which includes a large interactive media library of medical and surgical videos and animations. Users would then click on a part of the body to learn more about it. "The 3-D anatomical body is becoming a natural interface for contextualizing the library of information we already have," he says. Next, the body would be tailored to reflect the diseases and conditions of individual patients. It could even grow with you, he explains, "so you could go back to see your health at any previous point in time." Eventually, the map will also represent large populations and perhaps serve as an epidemiological tool. "What would the collective body of San Francisco or New York City look like?" Wong asks. □



PhineasMap floating above the Brooklyn Bridge. Courtesy of Virgil Wong, www.virgilwong.com/installations/phineasmap